Hardy & Bryman-3836-Ch-02:Hardy & Bryman-3836-Ch-02.qxp 12/12/2008 5:07 PM Page 15

PART I

Foundations

2

Constructing Variables

ALAN BRYMAN AND DUNCAN CRAMER

The process of quantitative research is frequently depicted as one in which theory is employed in order to deduce hypotheses which are then submitted to empirical scrutiny. Within the hypothesis will be two or more concepts that will require translation into empirical indicators. These indicators are frequently referred to as *variables* and represent the fundamental focus of all quantitative research. While some writers might question the degree to which quantitative research necessarily follows such a linear progression and indeed how far it is driven by hypotheses (as against simply research questions), there is no doubt that the variable represents a major focus (Bryman, 2001). It constitutes a crucial bridge between conceptualization and findings.

Essentially, the quantitative researcher is concerned to explore variation in observed values among units of analysis and the correlates and causes of variation. All techniques of quantitative data analysis – from the most basic methods to the most advanced – are concerned with capturing variation and with helping us to understand that variation. The variable is crucial because it is the axis along which variation is measured and thereby expressed. Indeed, so central is the variable to the discourse of quantitative research that it has to all intents and purposes become synonymous with the notion of a concept. Variables are, after all, supposed to be measures or indicators that are designed to quantify concepts, but frequently writers of research papers and methodology texts refer to the process of measuring variables. In the process, concepts and variables become almost indistinguishable. The variable is also frequently the focus of attention for critics of quantitative research (e.g., Blumer, 1956), in large part because it is emblematic of the research strategy.

The variable can be usefully contrasted with the idea of a *constant*. The latter occurs when there is no variation in observed values among units of analysis, as when all members of a survey sample reply to a questionnaire item in the same way. Uncovering constants is relatively unusual and is likely to require a somewhat different strategy on the part of the researcher, since techniques of quantitative data analysis are typically concerned with exploring variation rather than its absence.

LEVELS OF MEASUREMENT

One of the most fundamental issues in quantitative data analysis is knowing which types of technique can be used in relation to particular levels of measurement. It is fundamental because each statistical technique presumes that the levels of measurement to which it is being applied are of a certain type or at least meet certain basic preconditions. This means that if a technique is applied to variables which do not meet its underlying assumptions, the resulting calculation will be meaningless. Therefore, being able to distinguish between the different levels of measurement is basic to the art and craft of quantitative data analysis.

Writers often refer to different 'types of variables' as a shorthand for different levels of measurement. As such there is an array of different types of variables or levels of measurement. This array reflects the fact that the four levels of measurement to be discussed are on a continuum of degrees of refinement. There are four types of variables which are typically presented in terms of an ascending scale of refinement: nominal; ordinal; interval; and ratio.

Nominal variable

The *nominal variable*, often also referred to as the *categorical variable*, is the most basic level of measurement. It entails the arbitrary assignment of numbers (a process referred to as *coding*) to the different categories that make up a variable. The different categories simply constitute a classification. We cannot order them in any way – they are simply different. The numbers that are different have no mathematical significance; instead, they act as tags which facilitate the computer processing of the data. Thus, if we asked a question in a social survey on religious affiliation, we would assign a number to each type of affiliation and record each respondent's affiliation with the appropriate number. Similarly, in an experiment on asking questions, Schuman and Presser (1981) asked:

> The next question is on the subject of work. People look for different things in a job. Which of the following five things would you *most* prefer in a job?

The five options which could be chosen were:

1 Work that pays well
2 Work that gives a feeling of accomplishment
3 Work where there is not too much supervision and you make most decisions yourself
4 Work that is pleasant and where the other people are nice to work with
5 Work that is steady with little chance of being laid off

In assigning numbers to each of these five possible answers, all we are doing is supplying a label to each type of response. We can only say that all those answering in terms of the first response differ from those answering in terms of the second, who differ from those answering in terms of the third, and so on.

Sometimes, we have just two categories, such as male/female or pass/fail. Strictly speaking such variables – often referred to as *dichotomous variables* or *binary variables* (e.g., Bryman and Cramer, 2001) – are nominal variables. However, sometimes such variables require a different approach to analysis from nominal variables with more than two categories and are therefore treated by some writers as a separate type of variable.

Ordinal variable

As we have seen, with a nominal variable we can say no more than that people (or whatever the unit of analysis) differ in terms of its constituent categories. If we are able to array the categories in terms of rank order then we have an ordinal variable. Thus, if we asked a sample of people how satisfied they were with their jobs and presented them with the following possible responses, we would have an ordinal variable:

1 Very satisfied
2 Fairly satisfied
3 Neither satisfied nor dissatisfied
4 Fairly dissatisfied
5 Very dissatisfied

In this case, although the numbers attached to each category are merely used to allow the answers to be processed, we can say that each number has a significance that is *relative* to the others, since they are on a scale from 1 (denoting very satisfied) to 5 (denoting very dissatisfied). Each number therefore represents a level of job satisfaction or dissatisfaction. What we cannot say is that, for example, the difference between being very satisfied and fairly satisfied is the same as the difference between being very dissatisfied and fairly dissatisfied. All we can say is that the respondents differ in terms of their levels of job satisfaction, with some respondents being more satisfied than others.

Interval variable

An interval variable is the next highest level of refinement. It shares with an ordinal variable

Table 2.1 *Summary of the characteristics of the four types of variable*

	Is there a true zero point?	Are the distances between categories equal?	Can the categories be rank-ordered?
Ratio variable	Yes	Yes	Yes
Interval variable	No	Yes	Yes
Ordinal variable	No	No	Yes
Nominal variable	No	No	No

the quality of the rank ordering of the categories (which should more properly be called *values*) but differs in that with an interval variable, the distances between the categories are equal across the range of categories. Thus, we can say that the difference between a temperature of 43°F and 44°F is the same as the difference between 24°F and 25°F. As such, the values that an interval variable can take are genuine numbers rather than the scoring or coding process associated with the quantification of the categories of nominal and ordinal variables, where the number system is essentially arbitrary. However, interval variables are relatively unusual in the social sciences, in that most apparently interval variables are in fact ratio variables.

Ratio variable

A ratio variable represents the highest level of measurement. It is similar to an interval variable, but in addition there is a true zero point. In measurement theory, a true zero point implies an absence of the quality being measured, that is, you cannot have less than none of it. This feature means that not only can we say that the difference between an income of $30 000 a year and an income of $60 000 a year is the same as the difference between an income of $40 000 and an income of $70 000 a year (that is, a difference of $30 000), but also we can say that the income of $60 000 a year is double that of $30 000 a year. This means that we can conduct all four forms of arithmetic on ratio variables. Similar qualities can be discerned in such common variables as age, years in full-time education, size of firm, and so on.

In the social sciences, because most apparently interval variables are ratio variables, it is common for writers to prefer to refer to them as interval/ratio variables (e.g., Bryman and Cramer, 2001). Moreover, the vast majority of statistical techniques which require that the variable in question is at the interval level of measurement can also be used in relation to ratio variables. Therefore, the crucial distinctions for most purposes are between nominal, ordinal and interval/ratio variables.

Table 2.1 seeks to bring together the key decision-making principles that are involved in deciding how to distinguish between different kinds of variables.

MEASURES AND INDICATORS

A distinction is often drawn between measures and indicators. Measures constitute direct quantitative assessments of variables. For example, we could say that a question on respondents' incomes in a survey would provide us with a measure of the variable income. As such, reported income is a very direct estimate of income. This can be contrasted with a situation in which the quantitative assessment of a variable is or has to be indirect. An example is the previously cited question on job satisfaction. While the question asks directly about job satisfaction, we do not know whether it does in fact tap that underlying variable. In this case, we are using the question as an *indicator* of job satisfaction. Whether it does in fact reflect respondents' levels of job satisfaction is an issue to do with whether it is a *valid* indicator, about which more will be said below. The issue of whether something is an indicator or a measure is not to do with an inherent quality: if respondents' answers to a question on their incomes are employed as a proxy for social class, it becomes an indicator rather than a measure as in the previous illustration.

CODING

A key step in the preparation of data for processing by computer is *coding*. As has already been suggested in relation to nominal and ordinal variables, precisely because these variables are not inherently numerical, they must be transformed into quantities. Illustrations of the coding process have already been provided in relation to Schuman and Presser's (1981) question on work motivation and an imaginary example of a question on job satisfaction. In each case, the numbers chosen are arbitrary. They could just as easily start with zero, or the direction of the coding could be the other way around.

Coding in relation to social surveys arises mainly in relation to two kinds of situations. Firstly, in the course of designing a structured interview or self-administered questionnaire, researchers frequently employ *pre-coded questions*. Such questions include on the instrument itself both the categories from which respondents must choose and the code attached to each answer. Coding then becomes a process of designating on the completed questionnaires which code an answer denotes. The second kind of context arises in relation to the post-coding of open questions. Coding in this context requires that the researcher derives a comprehensive and mutually exclusive set of categories which can denote certain kinds of answer.

What is crucial is that the coding should be such that:

- the list of categories is mutually exclusive so that a code can only apply to one category;
- the list of categories is comprehensive, so that no category or categories have been obviously omitted; and
- whoever is responsible for coding has clear guidelines about how to attach codes so that their coding is consistent (often called *intra-coder reliability*) and so that where more than one person is involved in coding the people concerned are consistent with each other (*inter-coder reliability*).

The first two considerations are concerned with the design of pre-coded questions and with the derivation of categories from open questions. The third consideration points to the need to devise a coding frame which pinpoints the allocation of numbers to categories. In a sense, with pre-coded questions, the coding frame is incorporated into the research instrument. With open questions, the coding frame is crucial in ensuring that a complete list of categories is available and that the relevant codes are designated. In addition, it is likely to be necessary to include a detailed set of instructions for dealing with the uncertainties associated with the categorization of answers to open questions when the appropriate category is not immediately obvious. With techniques like structured observation and content analysis, the design of such instructions – which is often in a form known as a *coding manual* – is a crucial step in the coding of the unstructured data which are invariably the focus of these methods.

A further consideration is that researchers quite often *recode* portions of their data. This means that their analyses suggest that it is likely to be expedient or significant to aggregate some of the codes and hence the categories that the codes stand for. For example, in the coding of unstructured data, the researcher might categorize respondents into, for example, nine or ten categories. For the purposes of presenting a frequency table for that variable, this categorization may be revealing, but if the sample is not large, when a contingency table analysis is carried out (e.g. cross-tabulating the variable by age), the cell frequencies may be too small to provide a meaningful set of findings. In response to this situation, the researcher may group some of the categories of response so that there are just five categories. Such recoding of the data can only be carried out if the recoded categories can be meaningfully combined. There is the risk that the process of recoding in this way might result in combinations that cannot be theoretically justified, but recoding of data is quite common in the analysis of survey and other kinds of data.

SCALE CONSTRUCTION

One of the crucial issues faced in the measurement process in social research is whether to employ just one or more than one (and in fact usually several) indicators of a variable. Employing more than one indicator has the obvious disadvantage of being more costly and time-consuming than relying on one indicator. However, there are certain problems with a dependence on single indicators:

1. A single indicator may fail to capture the full breadth of the concept that it is standing in for. This means that important aspects of the concept are being overlooked. The use of more than one

indicator increases the breadth of the concept that is being measured.

2. In surveys, a single indicator may fail to capture a respondent's attitude to an issue or behaviour. This may be due to a variety of factors, such as lack of understanding or misinterpretation of a question. By using several indicators, the effect of such error may be at least partly offset by answers to other questions which serve as indicators and which are not subject to the same problem.
3. When more than one indicator is employed and the score on each indicator is then combined to form a total score for each respondent (as occurs with the use of summated scales – see below), much greater differentiation between respondents is feasible than when a single indicator is employed. For example, with the imaginary job satisfaction indicator used above, respondents could only be arrayed along a scale from 1 to 5. If more than one indicator is used and scores are aggregated, much finer quantitative distinctions become possible.

In other words, for any single respondent, reliance on a single indicator increases the likelihood of measurement error.

The recognition of the importance of multiple-indicator measures has resulted in a growing emphasis on the construction of scales. There are different approaches to scale construction, but most researchers employ *summated scales*, which entail the use of several items which are aggregated to form a score for each respondent. This allows much finer distinctions between respondents to be made (see point 3 above). One of the most common formats for this type of scale is the *Likert scale*, whereby respondents are presented with a series of statements to which they indicate their levels of agreement or disagreement.

To illustrate this approach to scale construction, consider an attempt by a researcher interested in consumerism to explore (among other issues) the notion of the 'shopaholic'. The following items might be used to form a Likert scale to measure shopaholicism:

1 I enjoy shopping.
Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

2 I look forward to going shopping.
Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

3 I shop whenever I have the opportunity.
Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

4 I avoid going shopping if I can.
Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

5 When I visit a town or city I don't know well, I always want to see the shops.
Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

6 Shopping is a chore that I have to put up with.
Strongly agree | Agree | Neither agree nor disagree | Disagree | Strongly disagree

Each reply will be scored. Various scoring mechanisms might be envisaged, but let us say that we want 5 to represent the highest level of shopaholicism represented by each answer and 1 the lowest, with 3 representing the neutral position. Notice that two of the items (4 and 6) are 'reverse items'. With the four others agreement implies a penchant for shopping. However, with items 4 and 6, agreement suggests a dislike of shopping. Thus, with items 1, 2, 3 and 5, the scoring from strongly agree to strongly disagree will go from 5 to 1, but with items 4 and 6 it will go from 1 to 5. This reversal of the direction of questioning is carried out because of the need to identify respondents who exhibit *response sets*, which have been defined as 'irrelevant but lawful sources of variance' (Webb et al., 1966: 19). An example of a response set to which Likert and similar scales are particularly prone is *yeasaying* or *naysaying*, whereby respondents consistently answer in the affirmative or negative to a battery of items apparently regardless of their content. Consequently, if a respondent answered strongly agree to all six items, we would probably take the view that he or she is not paying much attention to the content of the items, since the answers are highly inconsistent in their implications.

The scale would have a minimum score for any individual of 6 (presumably indicating a 'shopaphobe') and a maximum of 30 (a total 'shopaholic'). Most will be arrayed on the 23 points in between. A respondent scoring 5, 4, 4, 5, 3, 5, producing a score of 26, would be towards the shopaholic end of the continuum. A further feature of such scales is that essentially they produce ordinal variables. We cannot really say that the difference between a score of 12 and a score of 13 is equal to the difference between a score of 15 and a score of 16. However, most writers are prepared to treat such scales as interval/ratio variables on the grounds that the large number of categories (25 in this case) means that they approximate to a 'true' interval/ratio variable. Certainly, summated scales are routinely treated as though they are interval/ratio variables in journal papers reporting the results of research.

With a Likert scale, respondents indicate their degrees of agreement. While a five-point scale of agreement is employed in the above example, some researchers prefer to use seven-point scales (very strongly agree, strongly agree, agree, etc.) or even longer ones. Other types of response format for summated scales include the binary response format:

I enjoy shopping   Agree   Disagree

the numerical response format:

I enjoy shopping   5   4   3   2   1

(where 5 means Strongly agree and 1 means Strongly disagree)

and the bipolar numerical response format:

I enjoy shopping 7 6 5 4 3 2 1 I hate shopping

Once a scale has been devised and administered, the researcher needs to ask whether the resulting scale measures a single dimension. There are three highly related aspects to this question.

1. Is there an item (or are there items) showing a different pattern of response from those associated with the other constituent items? If there are, the offending item or items need to be eliminated from the scale. One way of checking for this possibility is to search out information on the *item–total correlations*. An inter-item correlation relates scores on each item to scores on the scale overall. If an inter-item correlation is much lower or higher than other inter-item correlations, it becomes a candidate for exclusion from the scale.
2. Is the scale internally reliable? This issue, which will be elaborated upon below, is concerned with the overall internal coherence of the items. Eliminating items which show a different pattern of response from the rest will enhance internal reliability.
3. Does the scale contain more than one dimension? If there are items which show a different pattern of response, it may be that there is a systematic quality to this variation such that the scale is not measuring a single dimension but possibly two or more. When this occurs, the nature of the underlying dimensions needs to be identified and named. Factor analysis is the most appropriate means of exploring this issue and will be given greater attention below.

The second of these aspects is concerned with the more general issue of the reliability of variables, which, along with validity, is a crucial issue in the evaluation of the adequacy of a variable.

RELIABILITY AND VALIDITY OF VARIABLES

Reliability and validity are crucial criteria in the evaluation of variables. In spite of the fact that these two terms are often used interchangeably in everyday speech, they refer to different aspects of the qualities of variables.

Reliability

Reliability is concerned with the consistency of a variable. There are two identifiable aspects of this issue: *external* and *internal reliability*. If a variable is externally reliable it does not fluctuate greatly over time; in other words, it is stable. This means that when we administer our scale of shopaholicism, we can take the view that the findings we obtain are likely to be the same as those we would find the following week. The most obvious examination of external reliability is to test for *test–retest reliability*. This means that sometime after we administer our scale, we readminister it and examine the degree to which respondents' replies are the same for the two sets of data. The chief difficulty with this method is that there are no guidelines about the passage of time that should elapse between the two waves of administration. If the passage of time is too great, test–retest reliability may simply be reflecting change due to intervening events or respondents' maturation. Furthermore, testing for test–retest reliability can become a major data collection exercise in its own right, especially when large samples are involved and when there are several variables to be tested.

Internal reliability is an issue that arises in connection with multiple-indicator variables. If a variable is internally reliable it is coherent. This means that all the constituent indicators are measuring the same thing. There are several methods for assessing internal reliability, one of which – item–total correlations – was briefly mentioned above. A further method is *split-half reliability*. This entails randomly dividing the items making up a scale into two halves and establishing how well the two halves correlate. A correlation below 0.8 would raise doubts about the internal coherence of the scale and perhaps prompt a search for low item–total correlations. In the case of the shopaholicism scale, the scale would be divided into two groups of three items, and respondents' scores on the two groups of items would be assessed. Nowadays, the most common method of estimating internal reliability is *Cronbach's alpha* (α), which is roughly equivalent to the average of all possible split-half reliability coefficients for a scale (Zeller and Carmines, 1980: 56). The usual formula is

$$\alpha = \frac{k}{k-1}\left(1 - \frac{1}{\sigma_x^2}\sum \sigma_i^2\right),$$

where k is the number of items; $\sum \sigma_i^2$ is the sum of the total variances of the items; and σ_x^2 is the variance of the total score (Pedhazur and Schmelkin, 1991: 93). If alpha comes out below 0.8, the reliability of the scale may need to be investigated further. Computer software programs such as SPSS include a facility whereby it is possible to request that the alpha for the scale be computed with a particular item deleted. If there is a sharp rise in the level of alpha when any item is deleted, that item will then become a candidate for exclusion from the scale.

An important consideration in the measurement process is that resulting variables will contain *measurement error* – variation that is separate from true variation in the sample concerned. Such measurement error is an artefact of the measurement instruments employed and their administration. For many researchers, assessing internal reliability is one way in which they can check on the degree of measurement error that exists in summated scales, although it cannot exhaust the range of possible manifestations of such error.

Validity

Validity is concerned with the issue of whether a variable really measures what it is supposed to measure. Can we be sure that our scale of shopaholicism is really to do with shopaholicism and not something else? At the very least, we should ensure that our scale exhibits *face validity*. This will entail a rigorous examination of the wording of the items and an examination of their correspondence with the theoretical literature on consumption. We might also submit our items to judges and invite them to comment on the wording of the items and on the goodness of fit between the items and what we might take shopaholicism to entail. However, face validity is only a first step in validity assessment.

C*riterion-related* validity assesses a scale in terms of a criterion in terms of which people are known to differ. This form of validity assessment can be viewed in terms of two forms. Firstly, testing for *concurrent validity* relates a variable to a contemporaneous criterion. Thus, we might ask respondents who are completing our shopaholicism scale how frequently they go shopping. If we found that there was no difference between shopaholics and shopaphobes in terms of the frequency with which they go shopping, we might question how well the scale is measuring the underlying concept. Equally, if the two types of shoppers clearly differ, our confidence is enhanced that the scale is measuring what it is supposed to be measuring. Secondly, testing for *predictive validity* relates a variable to a future criterion. Some months after we administer the shopaholicism scale we might recontact our respondents and ask them about the frequency with which they have been shopping in the previous month. Again, we would expect the shopaholicism scale to be able to discriminate between the frequent and occasional shoppers. Alternatively, we might ask our respondents to complete a structured diary in which they report the frequency with which they go shopping and the amounts of time spent on their expeditions.

Testing for *construct validity* entails an examination of the theoretical inferences that might be made about the underlying construct. It means that we would have to stipulate hypotheses concerning the construct (shopaholicism) and then test them. Drawing on theories about the consumer society and consumerism, we might anticipate that shopaholics will be more concerned about the sign value of goods than their use value. Consequently, we might expect they will be more concerned with the purchase of goods with designer labels. We could therefore design some questions concerned with respondents' predilection for designer brands and relate these to findings from our shopaholicism scale. Of course, the problem here is that if the theoretical reasoning is flawed, the association will not be forged and this is clearly not a product of any deficiencies with our scale.

These are the major forms of validity assessment. Other methods, such as *convergent validity*, whereby a different method is employed to measure the same concept, are employed relatively rarely because they constitute major projects in their own right.

One final point on this issue is that validity presupposes reliability. If you have an unreliable variable, it cannot be valid. If a variable is externally unreliable, it fluctuates over time and therefore cannot be providing a true indication of what it is supposed to be measuring. If it is internally unreliable, it is tapping more than one underlying concept and therefore is not a genuine measure of the concept in question.

DUMMY VARIABLES

One way of examining the association between a nominal or categorical variable (such as religious affiliation or nationality) and a non-nominal variable (such as income or life satisfaction) is to code the different categories of the categorical variable in a particular way called dummy coding (Cohen and Cohen, 1983). This procedure will be explained in terms of the following example. Suppose we wanted to determine the association between nationality and life satisfaction. To enable the relevant statistics to be computed, a small sample of fictitious data has been created and is presented in Table 2.2.

Table 2.2 *Life satisfaction in three nationalities*

	American	British	Canadian
	9	8	7
	7	5	7
	6	4	4
Mean	7.33	5.67	6.00

The categorical variable consists of three nationalities, American, British and Canadian. Each group consists of three people. The non-categorical variable comprises a 10-point measure of life satisfaction varying from 1 to 10, with higher scores representing greater life satisfaction. From the mean score for each nationality, we can see that the Americans have the greatest life satisfaction, followed by the Canadians and then the British. What we are interested in is not the association between particular nationalities and life satisfaction (e.g., being American and life satisfaction) but the association between the general variable reflecting these nationalities and life satisfaction (i.e., the association between nationality and life satisfaction).

The simplest way of expressing the association between the general variable of nationality and life satisfaction is in terms of the statistical coefficient called *eta squared*. Eta squared is the variance in life satisfaction attributed to the variable of nationality as a proportion of the total variance in life satisfaction. It can be worked out from an unrelated one-way analysis of variance. In this case eta squared is 0.194. This method does not involve dummy coding.

The dummy coding of a categorical variable may be used when we want to compare the proportion of variance attributed to that variable with the proportion of variance attributed to non-categorical variables (such as age) together with any other categorical variables (such as marital status). The method usually used to determine these proportions is multiple regression. Multiple regression can be represented by the following regression equation:

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

The dependent or criterion variable is often designated y and in our example is life satisfaction. The independent or predictor variables are usually signified by x_1 to x_k. One of the predictor variables in our example is nationality. Another predictor might be age. The contribution or weight of each predictor is normally the partial regression coefficient, which is generally symbolized as b_1 to b_k. The a is the intercept and may be referred to as the constant.

Multiple regression assumes that the predictor variables are dichotomous or noncategorical. Dichotomous variables (such as gender) have two categories (female and male) and may be treated as if they are noncategorical in that one category is arbitrarily assumed to be higher than another. For example, females may be coded 1 and males 2. This cannot be done with categorical variables having more than two categories because the numbers will be seen as reflecting an ordinal scale at the very least. For instance, if we coded Americans 1, Britons 2 and Canadians 3, multiple regression will assume that Americans have the highest value and Canadians the lowest, which might not be the case. We cannot order nationalities in terms of their mean score on life satisfaction (with Americans coded 1, Canadians 2 and Britons 3) because this order might not be the same for the other predictor variables (such as age). Consequently, we have to treat the categorical variable as if it were a series of dichotomous variables.

The simplest form of coding is *dummy coding*, where we assign a 1 to the units of analysis belonging to that category and 0 to units not belonging to that category. So, for example, we could code the three nationalities as shown in Table 2.3. Here we use one dummy variable to code all Americans as 1 and all non-Americans as 0. We use another dummy variable to code all Britons as 1 and non-Britons as 0. In this scheme Americans are represented by a 1 on the first dummy variable and a 0 on the second dummy variable. Britons are denoted by a 0 on the first dummy variable and a 1 on the second dummy variable. We do not need a third dummy variable to code Canadians because Canadians are represented by a 0 on both dummy variables. The category denoted by all 0s is sometimes known as the reference category. Thus, only two dummy variables are needed to represent these three categories.

The number of dummy variables required to code a categorical variable is always one less than the number of categories. So, if there are four categories, three dummy variables are necessary. It does not matter which category is denoted by 1s and 0s. In our example, Americans could have been coded 0 0, Britons 1 0 and Canadians 0 1. The results for the dummy variables taken together will be exactly the same. If the reference category is also coded in 1s and 0s, then one less than the total number of dummy variables will be entered into the multiple regression because one of them is redundant. The reference category is represented by the intercept a in the regression equation. So, the multiple regression equation for regressing the criterion of life satisfaction on the dummy coded categorical variable of nationality is:

Table 2.3 *Dummy variable coding of three nationalities*

Nationalities	d_1	d_2
American	1	0
British	0	1
Canadian	0	0

$$\underset{(y)}{\text{Life satisfaction}} = \underset{(a)}{\text{Canadian}} + \underset{(b_1x_1)}{b_1 \times \text{American}} + \underset{(b_2x_2)}{b_2 \times \text{British}}$$

The multiple correlation squared is 0.194, which is the same value as that for eta squared. Dummy coded variables representing a particular categorical variable need to be entered together in a single step in a hierarchical multiple regression analysis.

EFFECTS AND CONTRAST CODING

Two other ways of coding categorical variables are effects and contrast coding. Both these methods will explain exactly the same proportion of variance by the categorical variable as dummy coding. However, the partial regression coefficients may differ insofar as they represent different comparisons. If information on particular comparisons is also needed, the required comparisons have to be specified with the appropriate coding. With dummy coding, the constant is the reference category. In our example on nationality, the unstandardized partial regression coefficient for the first dummy variable essentially compares the mean life satisfaction of Americans with that of Canadians. Similarly,

Table 2.4 *Effects coding of three nationalities*

Nationality	e_1	e_2
American	1	0
British	0	1
Canadian	−1	−1

Table 2.5 *Contrast coding of three nationalities*

Nationality	c_1	c_2
American	1	−½
British	−1	−½
Canadian	0	1

the unstandardized partial regression coefficient for the second dummy variable compares the mean life satisfaction of Britons with that of Canadians. See Cohen and Cohen (1983) for further details.

With effects coding, the constant is the mean of all equally weighted group means, which is produced by coding one of the categories as −1 instead of 0, such as the Canadians as shown in Table 2.4. In this case, the unstandardized partial regression coefficient for the first effects-coded variable compares the mean life satisfaction of Americans with that of all three groups. The unstandardized partial regression coefficient for the second effects-coded variable contrasts the mean life satisfaction of Britons with that of the three nationalities.

Contrast coding enables other kinds of comparisons to be made provided that the comparisons are independent or orthogonal. As with dummy and effects coding, the number of comparisons is always one less than the number of groups. For example, if we wanted to compare Americans with Britons and Americans and Britons combined with Canadians, we would code the groups as indicated in Table 2.5. For the comparisons to be independent, the products of the codes for the new contrast-coded variables have to sum to zero, which they do in this case:

$$1 \times (-\tfrac{1}{2}) + (-1) \times (-\tfrac{1}{2}) + 0 \times 1 = -\tfrac{1}{2} + \tfrac{1}{2} + 0 = 0.$$

FACTOR ANALYSIS

Factor analysis is commonly used to determine the factorial validity of a measure assessed by several different indices. Factorial validity refers to the extent to which separate indices may be seen as assessing one or more constructs. Indices that measure the same construct are grouped together to form a factor. Suppose, for example, we were interested in determining whether people who said they were anxious were also more likely to report being depressed. We made up three questions for assessing anxiety (A1–A3) and three questions for measuring depression (D1–D3):

A1 I get tense easily
A2 I am often anxious
A3 I am generally relaxed

D1 I often feel depressed
D2 I am usually happy
D3 Life is generally dull

Each question is answered on a five-point Likert scale ranging from 'Strongly agree' (coded 1) through 'Neither agree nor disagree' (coded 3) to 'Strongly disagree' (coded 5).

The anxiety questions appear to ask about anxiety and the depression questions seem to be concerned with depression. If people can distinguish anxiety from depression and if people who are anxious tend not to be depressed as well, then answers to the anxiety questions should be more strongly related to each other than to the answers to the depression questions. Similarly, the answers to the depression questions should be more highly associated with each other than with the answers to the anxiety questions. If this turns out to be the case, the three items measuring anxiety may be combined together to form a single index of anxiety, while the three items assessing depression may be aggregated to create a single measure of depression. In other words, the anxiety items should form one factor and the depression items should form another factor.

However, the way the answers to these six questions are actually grouped together may differ from this pattern. At one extreme, each answer may be unrelated to any other answer so that the answers are not grouped together in any way. At the other extreme, all the answers may be related and grouped together, perhaps representing a measure of general distress. In between these two extremes the range of other possible patterns is large. For example, the two positively worded items (A3 and D2) may form one group of related

Table 2.6 *Coded answers on a 5-point scale to six questions*

Cases	A1(Tense)	A2 (Anxious)	A3 (Relaxed)	D1 (Depressed)	D2 (Happy)	D3 (Dull)
1	5	3	2	3	4	2
2	2	1	4	3	2	4
3	4	3	2	4	1	4
4	3	5	1	2	3	2
5	2	1	5	4	2	4
6	3	2	4	3	4	1

Table 2.7 *Triangular correlation matrix for six variables*

Variables	A1 (Tense)	A2 (Anxious)	A3 (Relaxed)	D1 (Depressed)	D2 (Happy)	D3 (Dull)
A1 (Tense)	1.00					
A2 (Anxious)	0.51	1.00				
A3 (Relaxed)	–0.66	–0.94	1.00			
D1 (Depressed)	–0.04	–0.61	0.51	1.00		
D2 (Happy)	0.33	0.22	–0.11	–0.59	1.00	
D3 (Dull)	–0.36	–0.45	0.29	0.63	–0.91	1.00

items and the remaining four negatively worded items may comprise another group of related items. We use factor analysis to see how the items group together.

Correlation matrix

The first step in looking at the way the answers are related to each other is to correlate each answer with every other answer. To illustrate our explanation we will use the small sample of fictitious data in Table 2.6. This table shows the coded answers of six people to the six questions on anxiety and depression. So, case number 1 answers 'strongly disagree' to the first question (A1) and 'neither agree nor disagree' to the second question (A2). Correlating the answers of the six cases to the six questions results in the triangular correlation matrix shown in Table 2.7.

Correlations can vary from –1 through 0 to 1. The sign of the correlation indicates the direction of the relationship between two variables. A negative correlation represents high scores on one variable (e.g., 5) being associated with low scores on the other variable (e.g., 1). For instance, from Table 2.7 we can see that the correlation between the answers to the questions about being anxious (A2) and being relaxed (A3) is –0.94. In other words, people who agree they are anxious have a strong tendency to disagree that they are relaxed (and vice versa). A positive correlation indicates high scores on one variable being associated with high scores on the other variable and low scores on one variable going together with low scores on the other variable. For example, in Table 2.7 we can see that the correlation between the answers to the questions about being tense (A1) and being anxious (A2) is 0.51. In other words, individuals who agree that they are tense have a moderate tendency to agree that they are anxious.

The strength of the association between two variables is indicated by its absolute value (i.e., disregarding the sign of the correlation). The correlation between being anxious and being relaxed (–0.94) is stronger than that between being tense and being anxious (0.51) because it is bigger. Conventionally, correlations in the range of 0.1 to 0.3 are usually described verbally as being weak, small or low; correlations in the range of 0.4 to 0.6 as being moderate or modest; and correlations in the range of 0.7 to 0.9 as being strong, large or high. The correlations in the diagonal of the matrix can be ignored or omitted as they represent the correlation of the variable with itself. This will always be 1.0 as there is a perfect positive relationship between two sets of the same scores.

From Table 2.7 it can be seen that the absolute size of the correlations among the three anxiety answers ranges from 0.51 to 0.94, suggesting that these answers go together. The absolute size of the correlations among the three depression answers ranges from 0.59 to 0.91, indicating that these answers go together. The data were deliberately generated to be associated in this way. In data that have not been so made up, the pattern may be less obvious. Even in these data, the pattern of results is not clear-cut. The absolute size of the correlation between being anxious (A2) and being depressed (D1) is 0.61, larger than the 0.51 between being tense (A1) and being anxious (A2). Furthermore, the correlation between being relaxed (A3) and being depressed (D1) is 0.51, the same as that between being tense (A1) and being anxious (A2). Consequently, it is possible that the answers to D1 may be more closely associated with the three anxiety items than with the other two depression items. Thus, the way the items are grouped may not be sufficiently apparent from simply looking at the correlations among the items. This is more likely to be the case the larger the number of variables. Factor analysis is used to make the way variables are grouped together more obvious.

Factor analysis is a set of statistical procedures that summarize the relationships between the original variables in terms of a smaller set of derived variables called factors. The relationship between the original variable and the factors is expressed in terms of a correlation or *loading*. The larger the absolute size of the correlation, the stronger the association between that variable and that factor. The meaning of a factor is inferred from the variables that correlate most highly with it. Originally, factor analysis was used to *explore* the way in which variables were grouped together. More recently, statistical techniques have been designed to determine whether the factors that have been obtained are similar to or *confirm* those that were either hypothesized as existing or actually found in another group. Consequently, when developing a series of indices to measure a variable, it may be more appropriate to use an exploratory rather than a confirmatory factor analytic technique. If we want to compare our results with those already obtained, then confirmatory factor analysis may be preferable.

Exploratory factor analysis

There are a number of different procedures for exploratory factor analysis. The two most commonly used are *principal components* and *principal factors* or *axes*. Factor analysis is the term used to describe all methods of analysis but may also refer to the particular technique called principal factors. In principal components all the variance in a variable is analysed. Variance is a measure of the extent to which the values of a variable differ from the mean. In principal components, this variance is set at 1.0 to indicate that all the variance in a variable is to be analysed. This will include any variance that may be due to error rather than to the variable being measured. In principal axes only the variance that the variable shares with all other variables in the analysis is analysed. This shared variance or covariance is known as *communality* and will be less than 1.0. Communality is also sometimes used to refer to the variance in principal components.

Often both procedures will give similar results so that it does not matter which procedure is selected. Tabachnick and Fidell (2001) have suggested that principal components should be used when an empirical summary of the data is required, whereas principal axes should be applied when testing a theoretical model. One problem with principal axes is that the communalities may not always be estimable or may be invalid (e.g., having values greater than 1 or less than 0), thereby requiring one or more variables to be dropped from the analysis. Consequently, we will use principal components to illustrate the explanation of factor analysis.

Initial factors The number of factors initially extracted in an analysis is always the same as the number of variables, as shown in Table 2.8. For each variable, the entries in the table represent its loading or correlation with each factor; the square of each entry is a

Table 2.8 *Initial principal components*

	1	2	3	4	5	6
A1 (Tense)	–0.62	0.39	0.67	–0.13	–0.02	0.00
A2 (Anxious)	–0.84	0.44	–0.23	0.20	0.09	0.00
A3 (Relaxed)	0.79	–0.61	0.10	0.05	0.03	0.00
D1 (Depressed)	0.77	0.26	0.54	0.21	0.04	0.00
D2 (Happy)	–0.69	–0.68	0.24	–0.08	0.11	0.00
D3 (Dull)	0.80	0.53	–0.15	–0.22	0.10	0.00
Eigenvalues	3.42	1.53	0.88	0.16	0.03	0.00
Eigenvalues as proportion of total variance	0.57	0.26	0.15	0.03	0.01	0.00

measure of variance. So, the variance of A1 is −0.62 squared, which is about 0.38. The amount of variance accounted for by a factor is called the *eigenvalue* or latent root, and is the sum of the squares of each entry in a column, that is, the sum of the variances for each variable. The first factor has the highest loadings and extracts or reflects the greatest amount of variance in the variables. It has an eigenvalue of 3.42. Subsequent factors represent decreasing amounts of variance. The second factor has an eigenvalue of 1.53, while the sixth factor has an eigenvalue of 0. The eigenvalues should sum to the number of factors, which in this case is 6 (allowing for rounding error). The variance that each factor accounts for can also be expressed as a proportion of the total variance. Thus, the first factor explains 3.42/6.00 = 0.57 of the total variance, and the second factor 1.53/6.00 = 0.26.

Number of factors to be retained Because the number of factors extracted is always the same as the number of variables that are analysed, we need some criterion for determining which of the smaller factors should be ignored as the bigger ones account for most of the variance. One of the main criteria used is the Kaiser or Kaiser–Guttman criterion, which was suggested by Guttman and adapted by Kaiser. This criterion ignores factors that have eigenvalues of 1 or less. The maximum variance that each variable explains is set at 1, so that factors having eigenvalues of 1 or less explain less variance than that of one variable on average. In other words, according to this criterion, only factors that account for the variance of more than one variable are retained for further analysis. In our example, only the first two factors have eigenvalues of more than 1, while the other four factors have eigenvalues of 1 or less. Thus, according to this criterion, we would keep the first two factors for further analysis. It should be noted that a cut-off at 1 may be somewhat arbitrary when there are factors which fall close to either side of this value. According to this criterion, a factor with an eigenvalue of 1.01 will be retained while one with an eigenvalue of 0.99 will be dropped, although the difference in the eigenvalues of these two factors is very small. In such cases it may be worthwhile extracting both more and fewer to see whether these factors, when rotated, are more meaningful than those retained according to Kaiser's criterion.

A second criterion is the graphical scree test proposed by Cattell (1966), who suggested that the Kaiser criterion may retain too many factors when there are many variables and too few factors when there are few variables. Child (1990) has specified 'many variables' as more than 50 and 'few' as less than 20. In the scree test the eigenvalue of each factor is represented by the vertical axis of the graph while the factors are arranged in order of decreasing size of eigenvalue along the horizontal axis, as shown in Figure 2.1.

Scree is a geological term for the rubble and boulders lying at the base of a steep slope and obscuring the real base of the slope itself. The number of factors to be extracted is indicated by the number of factors that appear to represent the line of the steep slope itself where the scree starts. The factors forming the slope are seen as being the substantial factors, while those comprising the scree are thought to be small error factors. The number

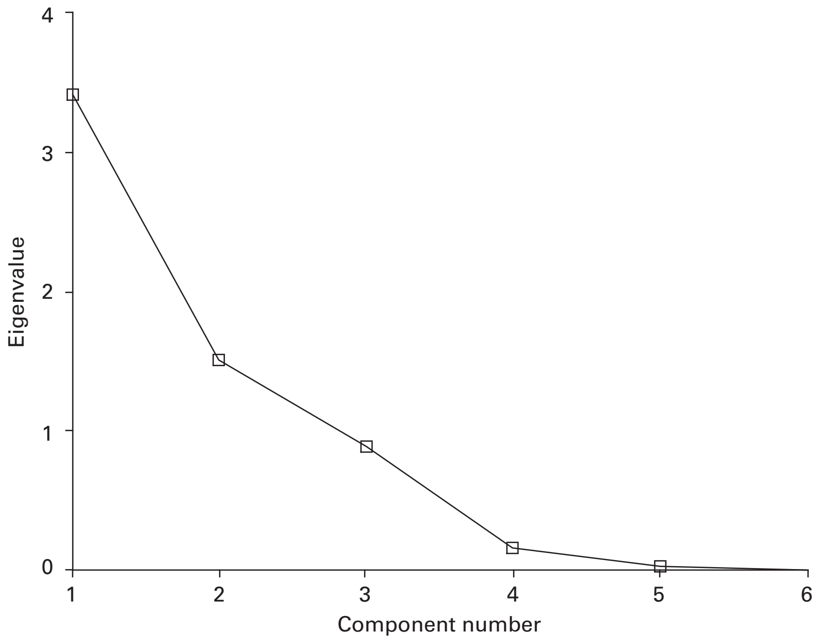

Figure 2.1 *Cattell's scree test*

of the factor identifying the start of the scree indicates the number of factors to be kept.

The scree factors are usually identified by being able to draw a straight line through or very close to their points on the graph. This is not always easy to do, as shown in Figure 2.1. In this case it is unclear whether the scree begins at factors 2, 3 or 4, and so whether the number of factors to be retained for further analysis should be 2, 3 or 4. Thus, one problem with the scree test is that determining where the scree begins may be subjective, as in this example. When this occurs, it may be useful to extract both fewer and more factors around the number suggested by the scree test and to compare their meaningfulness when rotated. If more than one scree can be identified using straight lines, the number of factors to be retained is minimized by selecting the uppermost scree.

Factor rotation As already explained, the first factor in a factor analysis is designed to represent the largest amount of variance in the variables. In other words, most of the variables will load or correlate most highly with the first factor. If we look at the absolute loadings of the variables on the first factor in Table 2.8, we see that they vary from 0.62 to 0.84. The second factor will reflect the next largest amount of variance. As a consequence, the loadings of the variables on the second factor will generally be lower. We see in Table 2.8 that they range in absolute value from 0.26 to 0.68. The loading of variables on two factors can be plotted on two axes representing those factors, as shown in Figure 2.2. These axes are called reference axes. In Figure 2.2 the horizontal axis represents the first factor and the vertical axis the second factor. The scale on the axes indicates the factor loadings and varies in steps of 0.2 from –1.0 to +1.0. The item on anxiousness (A2), for example, has a loading of –0.84 on the first factor and of 0.44 on the second (see Table 2.8).

It may be apparent that the two axes do not run as close as they could to the points representing the variables. If we were to rotate the axes around their origin, then these two axes could be made to pass nearer to these points, as shown in Figures 2.3 and 2.4.

The effect of rotating the axes is generally to increase the loading of a variable on one of the factors and to decrease it on the others, thereby making the factors easier to interpret. For example, in Table 2.9 we can see that the effect of rotating the two axes is to increase the loading of the item on anxiousness from –0.84 to –0.91 on the first rotated

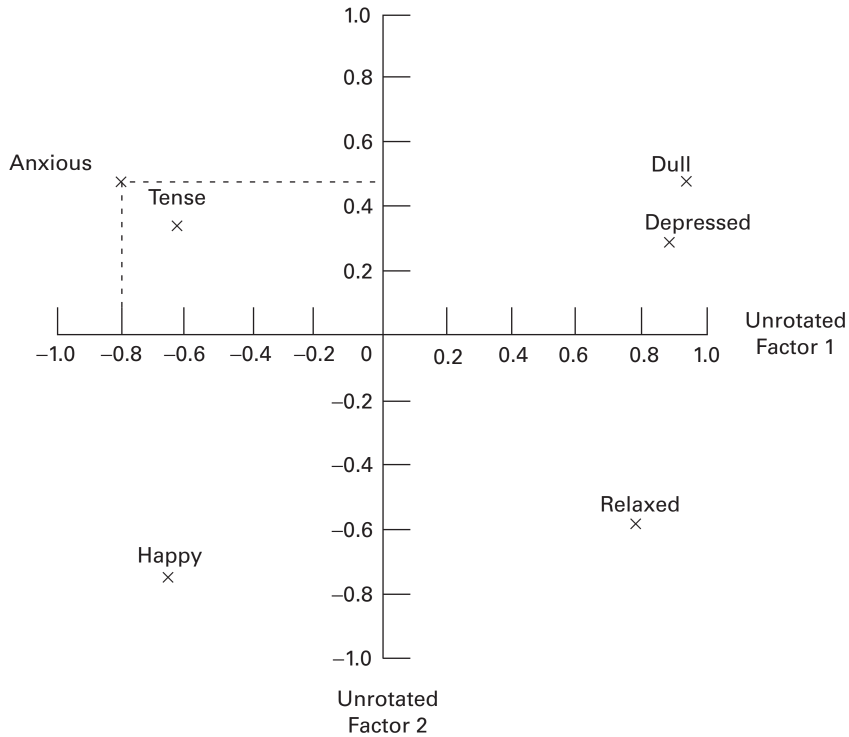

Figure 2.2 *Plotting variables on two unrotated factors*

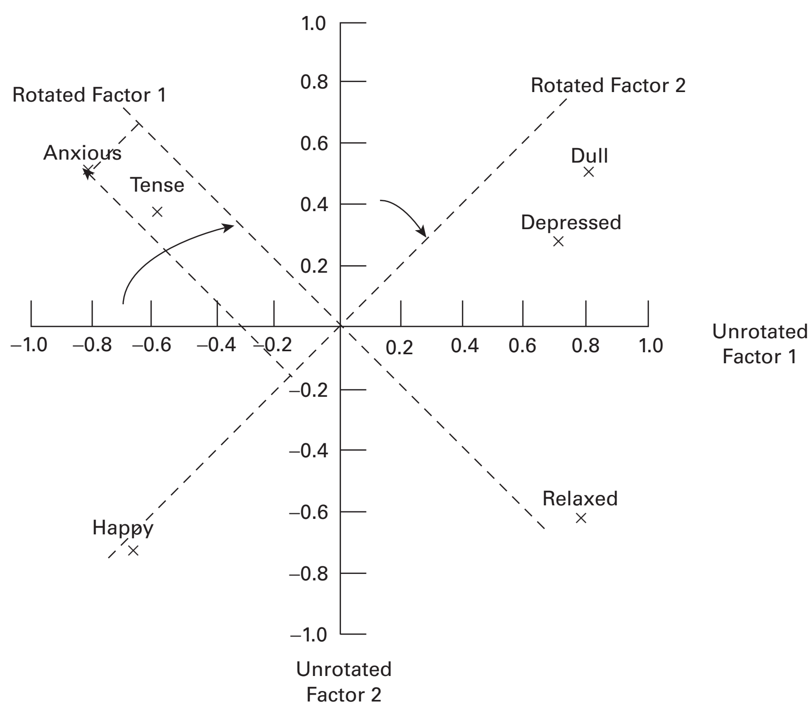

Figure 2.3 *Initial factors orthogonally rotated*

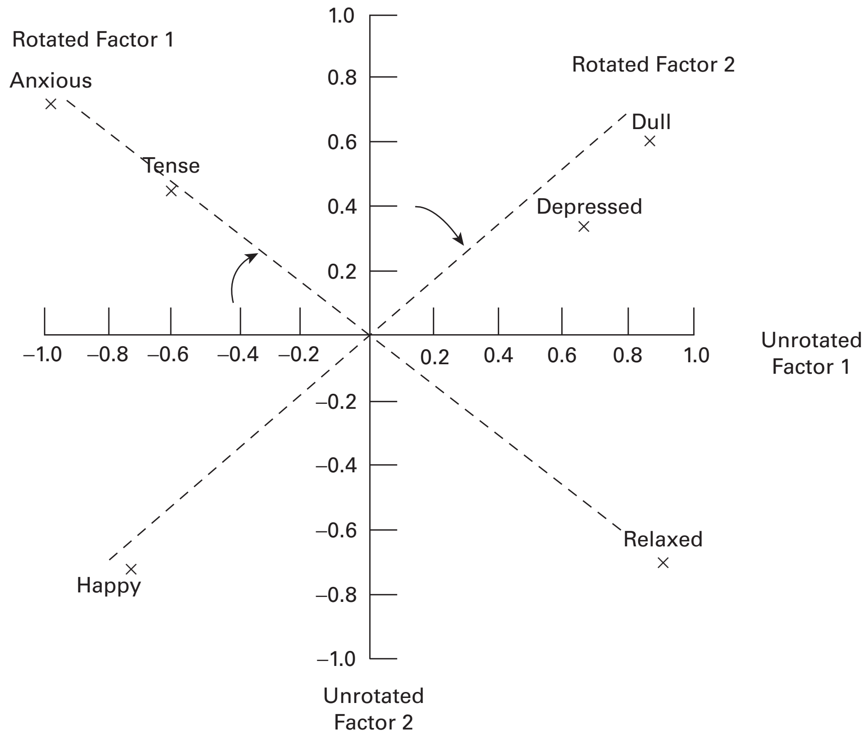

Figure 2.4 *Initial factors obliquely rotated*

Table 2.9 *First two orthogonally rotated principal components*

	1	2
A1 (Tense)	–0.72	–0.15
A2 (Anxious)	–0.91	–0.27
A3 (Relaxed)	0.99	0.11
D1 (Depressed)	0.37	0.72
D2 (Happy)	0.03	–0.96
D3 (Dull)	0.22	0.94
Eigenvalues	2.51	2.43
Eigenvalues as proportion of total variance	0.42	0.41

factor and to decrease it from 0.44 to 0.27 on the second rotated factor.

Axes may be rotated in one of two ways. First, they may be made to remain at right angles to each other, as is the case in Figure 2.3. This is known as *orthogonal* rotation. The factors are independent of or uncorrelated with one another. The advantage of this approach is that the information provided by the factors is not redundant. Knowing the values on one factor (e.g., anxiety) does not enable one to predict the values of another factor (e.g., depression) as the factors are unrelated. The disadvantage is that the factors may be related to one another in reality and so the factor structure does not accurately represent what occurs.

Second, the factors may be allowed to be related and to vary from being at right angles to one another, as illustrated in Figure 2.4. This is known as *oblique* rotation. The advantage of this method is that the factors may more accurately reflect what occurs in real life. The disadvantage is that if the factors are related, knowledge about the values of one factor may allow one to predict the values of other factors. The results of the two methods may be similar, as in this example.

The most widely used form of orthogonal rotation is *varimax*, which maximizes the variance within a factor by increasing high loadings and decreasing low loadings. The loadings shown in Table 2.9 were derived using this method. Comparing the results of Tables 2.8 and 2.9, we can see that orthogonal rotation has increased the loadings of three variables for the first (A1, A2 and A3) and second factor (D1, D2 and D3). It has decreased the loadings of three variables for the first (D1, D2 and D3) and second factor (A1, A2 and A3). The variables loading highest on the first factor are being relaxed (0.99), not anxious (–0.91) and not tense (–0.72) respectively, indicating that this factor represents anxiety. The variables loading highest on the second factor are not being happy (–0.96), finding life dull (0.94) and being depressed (0.72) respectively, showing that this factor reflects depression. These results suggest that the three items on anxiety (A1, A2 and A3) can be aggregated to measure anxiety and the three items on depression (D1, D2 and D3) can be grouped together to assess depression. Orthogonal rotation also has the effect of spreading the variance across the factors more equally. The variance accounted for by the first factor is 0.57 when unrotated and 0.42 (2.51/6.00) when rotated. For the second factor it is 0.26 when unrotated and 0.41 (2.43/6.00) when rotated.

The results of an oblique rotation using a method called *direct oblimin* are presented in Table 2.10. The findings are similar to those for varimax. The variables loading highest on the first factor are being relaxed (0.99), not anxious (–0.94) and not tense (–0.73), respectively. The variables loading highest on the second factor are finding life dull (0.96), not being happy (–0.95) and being depressed (0.78), respectively. The results indicate that the three anxiety items (A1, A2 and A3) can be combined together, as can the three depression items (D1, D2 and D3). The two factors were found to have a correlation of 0.36 with one another. As negative values on the first factor indicate anxiety and positive values on the second factor depression, the positive correlation between the two factors means that depression is associated with low anxiety. Because the factors are correlated, the proportion of variance explained by each factor cannot be estimated as it is shared between the factors.

Combining items to form indices The results of the factor analysis are used to determine which items should be combined to form the scale for measuring a particular construct. Items loading highly on the relevant factor (e.g., anxiety) and not on the other factors (e.g., depression) should be used to form the scale. The direction of scoring for the scale needs to be established. Generally higher scores on the scale should indicate greater quantities of the variable being measured. For example, if the scale is assessing anxiety, it is less confusing if high scores are used to denote high anxiety rather than low anxiety. The numerical codes for the responses may have to be reversed to reflect this. For instance, the numerical codes for the anxiety items A1 and A2 need to be reversed so that strong agreement with these items is recoded as 5. The scale should have adequate alpha reliability. Items not contributing to this should be omitted.

Table 2.10 *First two obliquely rotated principal components*

	1	2
A1 (Tense)	–0.73	–0.28
A2 (Anxious)	–0.94	–0.43
A3 (Relaxed)	0.99	0.29
D1 (Depressed)	0.50	0.78
D2 (Happy)	–0.20	–0.95
D3 (Dull)	0.39	0.96

CONCLUSION

In this chapter, we have moved fairly rapidly from some very basic ideas concerning variables to some fairly complex approaches to their creation and assessment. However, in another sense, the entire chapter deals with issues that are fundamental to the analysis of quantitative data, since the variable is the basic reference point. We have explored several ways in which variables are created, both in terms of such strategies as summated scales, which are common in the measurement of attitudes, and in terms of the ways in which analysts seek to refine and improve the quality of variables. Since the variable is fundamental to all quantitative data analysis, the material covered in this chapter constitutes an important starting point for many of the chapters in this book that deal with various aspects of quantitative data analysis.

REFERENCES

Blumer, H. (1956) 'Sociological analysis and the "variable"', *American Sociological Review*, 21, 683–90.

Bryman, A. (2001) *Social Research Methods*. Oxford: Oxford University Press.

Bryman, A. and Cramer, D. (2001) *Quantitative Data Analysis with SPSS Release 10 for Windows: A Guide for Social Scientists*. London: Routledge.

Cattell, R.B. (1966) 'The scree test for the number of factors', *Multivariate Behavioral Research*, 1, 245–76.

Child, D. (1990) *The Essentials of Factor Analysis* (2nd edition). London: Cassell.

Cohen, J. and Cohen, P. (1983) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd edition). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pedhazur, E.J. and Schmelkin, L.P. (1991) *Measurement, Design and Analysis: An Integrated Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schuman, H. and Presser, S. (1981) *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*. San Diego, CA: Academic Press.

Tabachnick, B.G. and Fidell, L.S. (2001) *Using Multivariate Statistics* (4th edition). New York: HarperCollins.

Webb, E.J., Campbell, D.T., Schwartz, R.D. and Sechrest, L. (1966) *Unobtrusive Measures: Nonreactive Measures in the Social Sciences*. Chicago: Rand McNally.

Zeller, R.A. and Carmines, E.G. (1980) *Measurement in the Social Sciences: The Link between Theory and Data*. Cambridge: Cambridge University Press.