Oppenheim, A.N. Questionnaire Design, Interviewing and Attitude Measurem

London: Pinter
1992

9 SOME BASIC MEASUREMENT THEORY

Non-factual questions

As we have seen towards the end of the preceding chapter, there are serious objections to the use of single questions to measure such non-factual subjects as awareness, percepts, social representations, brand images, opinions, beliefs, attitudes, values and stereotypes. Such issues are usually more complex than questions of fact; they have to do with states of mind, rather than with behaviour or with events in the outside world, and are therefore difficult to measure and to validate; they are generally multi-faceted and have to be approached from several directions; above all, single questions dealing with such sensitive topics are much more open to bias and unreliability due to wording, question format and contextual effects. It is not unusual to find that responses by the same group of subjects to two virtually identical items differ by as much as 20 per cent or more, and we have no guiding principles by which to predict how respondents will react to various forms of wording of what is basically the same question.

Opinion researchers who commonly rely on single attitudinal questions (say, whether or not the prime minister is doing 'a good job') are taking considerable risks. However, they would argue that these risks can be minimized by using the same question in successive surveys, so that an analysis of trends-over-time becomes possible — whatever the question 'means' to respondents. They might also argue that, since they have used this question many times in other surveys, they know a good deal about its properties and correlates with other variables. Also, something depends on what we want to know: perhaps all we need is a quick popularity snapshot of the prime minister, rather than an in-depth analysis of what people mean by 'a good job' and how this percept has been influenced by recent events.

But to the researcher who is hoping to rely on single questions about people's attitudes to their doctor, to animal rights, to immigrants, to Green policies or to a certain brand of washing-up liquid, the unreliability of such single questions will pose a serious problem. As we have indicated, if these variables are at all important to the study, then the way forward may well be through the multiple-question or *scaling* approach. To explain what is meant by scaling, we shall first have to consider some aspects of basic measurement theory; after that we shall look at the linear scaling model in its classical form and then see how this model has been applied to the field of opinion and attitude measurement, and to the measurement of concepts, awareness and other subjective variables.

The linear scaling model

WHAT IS A MEASURE?

Let us start with a simple example. Suppose that we wish to know the length of a table; we may take a metal foot-rule, hold it against the longer edge of the table, and read off the answer: say, thirty-eight inches. This measure has, in principle, two components: the 'true' measure and a second component which we may call 'error'. We shall, for the moment, ignore the kinds of errors-of-observation and notation that may arise because we have been a little careless or imprecise, and draw attention to the fact that we have used a *metal* foot-rule. Since we know that metal expands or contracts under different temperature conditions, part of our answer is not a measure of length but a measure of temperature. Since, at this point, we are not interested in temperature, we regard this aspect of the measure as 'error' — in the sense that it is irrelevant to what we need.

In this diagrammatic form, any measure could be represented as shown in Figure 9.1.

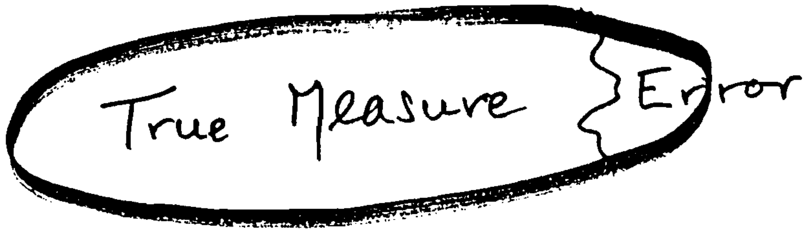

Figure 9.1.

The oval shape represents the reading or observation; the true measure is that part of the reading which we want — it is an indicator of an attribute of something or someone, for example length, musical ability, blood pressure; error is that part of the measure which we don't want. The wavy line between the two parts indicates that we are rarely certain about the exact size or location of these components. The diagram is an abstraction; it shows that we know our measure is 'impure' or 'contaminated' but not always by how much or in what way.

EARLY SCALES IN THE NATURAL SCIENCES

You might argue at this point that the degree of contraction or expansion in a metal foot-rule under normal everyday conditions is very slight and that consequently the amount of error introduced in our measure is negligible — or that we should have used a wooden ruler in the first place! The issue which this example illustrates is, however, a general one which is present in every kind of science.

Let us, by way of another familiar example, look at the ordinary household or medical thermometer. Once again, the measure which we seek is an abstraction: an indicator or an attribute of something or someone. 'Temperature', as such, is an abstraction, a scientific concept invented by humans; we try to measure it by observing the rise or fall of a column of mercury in a glass tube. We cannot say that the position of the mercury *is* the temperature; the mercury is only mercury. We are interested in the position of the mercury in the tube because we have made certain arrangements which will enable us to use it as an indicator of the state of someone (for example, a patient) or something (for example, the air in a greenhouse). These 'arrangements' are important. Some of them have been designed to minimize the kind of error-component which we discussed above. For example, if there were air in the tube above the mercury, then this would make it harder for the mercury to rise, the more so as the mercury nears the top of the tube, and so we would have been measuring both temperature and — to an increasing degree — air pressure. This error would, in other words, have taken the form of a systematic or correlated *bias* because the error gets greater as the readings increase, and vice versa. The error is therefore confounded, rather than uncorrelated or random.

By drawing the air out of the tube above the mercury we usually manage to eliminate this error to a large extent but not altogether. We note further that the mercury moves in a tube made of glass; as the temperature rises, the glass expands, the tube widens and the mercury rises more slowly — another 'confounded' error. We trust and assume that the walls of the tube are smooth and straight on the inside, for if they were uneven this would introduce a random error, and if they were, say, closer together at the top of the tube than at the bottom, we would have another 'confounded' error — though one that works in the opposite direction — yielding an over-estimate rather than an under-estimate.

Another important arrangement with thermometers is that they have a *scale*. In the eighteenth century, when the thermometer was being invented and improved by men such as Celsius, Réaumur and Fahrenheit, it seemed a good idea to think of temperature as a straight line, a linear dimension, and to divide this line into equal intervals (grouped according to the decimal system). The idea of using equal units along a continuum probably came from measures of distance such as miles or leagues, or perhaps from ells or yards or from measures of weight such as pounds or kilograms. However, two difficulties arose: one was the question of the size of the unit or interval, the other was the problem of finding a suitable starting point. After all, with distance, or length, or weight, a zero- or nil-point at the outset seems obvious — but can one think of any point as 'nil temperature', as no temperature at all? There was also the desire to invent something that could be used in a practical way both in the laboratory and in everyday life, covering the *range* of temperatures that were of interest at that time — the idea of measuring the temperature of the sun must have seemed quite out of the question! And so each of the inventors of the thermometer chose a zero-point and a top-of-the-scale point that seemed convenient for range and were thought to be 'anchored in Nature', be it the freezing and boiling points of water, or the temperature of the average healthy adult human body. Fahrenheit actually started his scale at the lowest point he could reach by using a mixture of ice, water and sal ammoniac or salt — this must have seemed like 'absolute zero' at the time. Celsius originally proposed that the steaming point of water should be zero and the freezing point 100°! Réaumur used a spirit thermometer, not a mercury one; instead of taking two chosen points and dividing the distance between them into 100 equal intervals, he considered the rate of expansion of the original volume of spirit as the temperature rose. He defined 1° as an expansion of one-thousandth of the original spirit volume; this led to a scale of eighty degrees or intervals between the freezing and boiling points of water. Today, in the physics laboratory, these early primitive methods of thermometry have been left far behind, but they serve to illustrate the basic problems of measurement that all sciences have to overcome.

We can ask, for example: is it right to think of changes in temperature as lying 'in a straight line'? This is the question of *linearity*, and the answer will lie in a better scientific understanding of the molecular processes that produce changes in temperature in different materials. We can also ask: is it possible to measure 'pure temperature' and nothing else, not even the expansion of the glass? This is the question of *uni-dimensionality* and is closely linked to the earlier notions of error. We know now that water boils at different temperatures at different heights, in accordance with barometric pressures. This, and many similar problems, such as the expansion rate of mercury or of certain gases, raises the problem of *standardization*: how can we be sure that any two or more thermometers will show the same readings under the same conditions? Then there is the question of consistency: can we rely on our thermometer to perform over and over again in the same way, yielding *reliable* readings which we can trust?

Another complex issue is the question of the *units of measurement*. As we have seen, three different inventors produced degrees of different sizes, which may cause us perhaps nothing worse than some irritating conversion problems, but we may ask whether at certain points — say, when something melts, or freezes, or burns, or dies — are there not also *qualitative*, as well as quantitative, differences in temperature? Or, to put it another way, are the units really interchangeable all along the scale? Are, say, the five degrees from 0° to 5°C (which cover thawing) the 'same' as the five degrees from 95° to 100°C (which cover the boiling of water at sea-level pressure)? And what about the beginning and the end, the top and bottom points? Do all measures require a *zero-point*? A maximum? Can they have negative readings? Last but not least, how can we be sure that there is a precise and consistent relationship between the abstract quality we call 'temperature' and the readings we obtain from mercury moving in a capillary tube? This is the problem of *validity* to which we shall have repeated occasion to return.

To sum up then: one answer to the question 'what is a measure?' is to describe the properties of a scale. A scale must satisfy the conditions of linearity and uni-dimensionality; it must be reliable (consistent); it must have units of measurement (not necessarily equal ones); these units will, if possible, be interchangeable; the scale needs some 'anchoring points' obtained through standardization or in relation to fixed, observable events, to give meaning to the scores; and most particularly, the scale must be valid, in the sense that it measures the attribute it sets out to measure. Perhaps surprisingly, the simple yardstick did satisfy all these requirements, once it was standardized (such a standard can still be viewed on the outside wall of the Greenwich observatory).

The basic notions of measurement are not hard to grasp, but they become more difficult to achieve as we move from the natural sciences, through the biological and medical sciences, to the behavioural and social sciences.

THE PROPERTIES OF A SCALE OF INTELLIGENCE

In the light of these requirements, let us look at a test of intelligence. Will it satisfy the conditions of linearity and uni-dimensionality? Some complex statistical techniques will have to be used to establish this for the particular scale in question, for we are no longer dealing with a simple foot-rule. (Even if uni-dimensionality can be established, how happy are we to think of a complex set of qualities, such as intelligence, in terms of a straight line?) Will it be reliable, in the sense that repeated administration of the test will yield virtually the same results? Leaving practical problems aside, the answer will probably be 'Yes', since we have fairly accurate techniques for assessing the degree of reliability. Will it have units of measurement, and, if so, will they be equal and interchangeable? Here we are on more difficult ground. The test is likely to yield a score in terms of points, which are treated as if they are units of equal interval which are interchangeable in many cases, but this remains a questionable issue.

Can the scores be given meaning, by relating the individual's score to some anchoring points and/or standardized ranges? To this, the answer will be 'Yes' — any test will be published together with its norms for various age groups, ability groups etc., and frequently the test scores will be converted into an Intelligent Quotient which offers, not a zero-point but a mid-point, the 'average' of 100 IQ. It is, however, highly debatable whether IQ points or intervals are equal units and whether they are interchangeable, though they are frequently treated as such.

Finally, will the test be valid? Will it really measure 'intelligence'? The first and crude answer to this will be 'Yes' because the test will have been correlated with some earlier, well-established IQ tests. But this merely moves the problem a stage further back, for how do we know whether *those* tests, the 'criterion measures', are valid? At this point we might have to enter two complex sets of arguments, one having to do with techniques of validation and one with the nature of intelligence. We may note, therefore, that even such a well-tried-out technique as a modern test of intelligence may — from the measurement point of view — have some rather dubious attributes.

ARE SCALES APPROPRIATE TO OUR DOMAIN?

Of course, it is also possible, indeed desirable, to turn the argument on its head. Instead of asking whether tests of intelligence satisfy the criteria of linear scaling, we might well ask whether the linear-scaling approach is appropriate to the complex set of qualities we call 'intelligence', or can ever hope to do it justice? More broadly, to what extent are the methods of measurement which have been so successful in the physical sciences still appropriate when we have to deal with behavioural and social phenomena? This remains a very controversial question on which the last word will not be spoken for a long time to come.

But if the linear-scaling approach makes demands which are technically hard to satisfy, are there any practicable alternatives? There are many positive answers to this question, but all of them entail losses as well as gains. One type of loss has to do with the richness of human experience. When we describe a rainbow in terms of a spectrum of lightwave frequencies, or motherlove in terms of a score on an attitude scale, we become acutely aware that something important is lost. On the other hand, the richer, more subjective approaches — for example, historical descriptions, diaries, literary accounts, free-style interviews, school reports, group discussions — are less reliable, less valid, less objective and less comparable. In the strict sense of the term they are not data or measures at all. They are rather like statements about hot or cold made by people before the thermometer was invented: subjective, imprecise, not always consistent and with little hope of consensus or comparability. The movement towards making things more 'objective' and 'scientific' is an attempt to get away from subjectivity, to make measures more 'true', more reliable, more precise, more replicable and more comparable even if this does entail certain losses. But it has to be said that the present 'state of the art' in measurement in the human sciences is often so crude and so poorly developed that the losses may well outweigh the gains. It thus becomes a key responsibility of competent research workers to choose the data-generation and -collection techniques which are most appropriate and most likely to 'do justice' to their problem and their subject-matter, a bitter and difficult choice between too much over-simplification for the sake of 'good measurement' and a subtler, more flexible approach which may not yield valid, replicable results. We can only hope that increases in the diversity and sophistication of approaches to measurement will eventually release us from this dilemma.

Other types of measures

ORDINAL SCALES

Let us now look at some well-understood departures from the linear scaling model. We have previously seen that, ideally, such a scale should consist of identical, interchangeable units. If at certain points along the continuum we need greater precision or accuracy, we can sub-divide such units and obtain finer-grained readings, but this does not alter the basic principle. However, what if the units are known to be unequal or cannot be shown to be equal? In such cases we look to see if they have a *sequence*, or an ordinal property — in other words, whether they can be *ranked*. In everyday life we quite often use a ranking approach: children may be ranked from top to bottom in a class; athletes, or horses, may be ranked in the order in which they finish a race; the pop charts may rank music recordings in order of their popularity. The important point to note is that an ordinal or ranking scale tells us nothing about the intervals between its points: the winning horse may have been ahead by no more than a neck, whereas the third one could have been trailing by several lengths. So long as we are certain about the ordinal properties of the scale, we can make use of these for the purposes of measurement. There are statistical techniques that will enable us to do this, and ordinal scales — while necessarily less precise than equal-interval ones — should otherwise have most of the other properties of linear scales.

A special sub-variant of ranking scales are scales using the *paired comparisons* method. This can be useful when we have a small number of objects (say, a dozen) and we wish to see how or in what order they should be arranged. For example, we might have a set of pictures, animals, smells or holiday resorts to place in order of preference, if we can. Of course, we might ask each of our respondents to rank the stimulus objects, but a more general method would be to arrange the objects in all possible pairs and present these to each respondent in random order, pair by pair. Instead of having to arrange a dozen pictures, or holiday resorts, in order of preference (which may be very difficult), all that the respondent has to do is to choose between two members of each pair. If there *is* some kind of underlying dimension or continuum on which all the objects can be placed, this will emerge from the statistical analysis, but even so, it will yield not an interval scale, but a rank-order scale as above. This method offers advantages where the dimension is uncertain and ranking would be difficult for the entire set of objects, but with human respondents it can only be used with relatively small sets of stimuli — even twelve objects yield sixty-six paired comparisons to make, and the numbers very soon rise to unmanageable magnitudes.

Nominal measures

We now come to the measurement of nominal or categorical data. These are data which have no underlying continuum, no units or intervals which have equal or ordinal properties, and consequently cannot be scaled. Examples might be: eye colour; the daily newspaper you read; the political party you last voted for; the type of novel you like to read; ownership of consumer durables, such as a car, a home freezer, a motor mower; the brand of cigarette you smoke; your country of birth; and so on. There is no underlying linear scale; instead, there are a number of discrete categories into which the responses can be classified or 'coded', but the categories can be placed in any order and they have no numerical value, nor is there assumed to be an underlying continuum. Essentially, the only possible score for each category is a binary one: yes/no, or present/absent. Thus, if my eyes are brown and I am given a list of eye colours, then my answers will be 'no', 'no', 'no', until I come to 'brown' when I shall give a 'yes' response, and then 'no', 'no', 'no' again to the remainder of the list. There is no order to the categories since no colour is 'better' or 'stronger' or 'bigger' than any other.

Other measures

Beyond this, and in a sense departing still further from the linear-scaling model, are numerous other techniques using multiple items; for example, *perceptual techniques* such as the semantic differential; *sociometry*; various *grid techniques*; and *projective techniques*. These will be discussed in subsequent chapters.

Statistical treatment

Any accumulation of measures will require some form of management and interpretation, even if this is done in a purely intuitive way. More commonly, the data are subjected to statistical analysis, aided by computers. Typically, the measures for each respondent will be entered into a computer (or, in the case of small-scale manual analysis, on to an analysis sheet laid out in the form of a grid), and thereafter they will be manipulated in a variety of ways. We may calculate various averages, we may compute percentages, we could examine the data for statistical significance, we could work out correlations and we might go on to various forms of multi-variate analysis such as multiple regression or factor analysis. This should enable us to 'make sense of the data'; that is, to test our hypotheses, to compare results for various sub-groups, and so on.

It is very important to realise that the statistical treatment of linear interval scales is quite different from the treatment of nominal or categorical measures. The properties of the linear scale permit us to treat the scores as integers which may be added, subtracted, divided, multiplied and so on. This implies that they can be analysed by means of statistical techniques applicable to interval-type scales. These techniques are quite powerful; they include the familiar 'average' or mean, the variance and the standard deviation, analysis of variance, many types of correlation coefficient and most multi-variate methods. The results can be tested for statistical significance (to check whether the observed findings could have arisen by chance) by means of t-tests, F-tests, Z-tests and others, which many computer programs provide as a matter of course.

Not so with nominal data. Since they lack interval properties and have no underlying continuum, they cannot be added, multiplied or squared. In the initial stages, after entering our data, all we can do is count. For each of our categories (of eye colour, party affiliation or whatever), we shall wish to see *how many* of our respondents endorsed it: how many people in our sample had blue eyes, how many had brown, how many had grey, and so on. No numerical value can be attached to these counts or entries; all we can do is count the frequency of their occurrence. After that, we might wish to compare sub-samples: do men have grey eyes more often than women? Are older respondents more prone to vote for party P than younger ones? To make allowances for differences in sub-sample sizes, we shall wish to convert these frequency distributions into *percentages*, and then study the differences between the groups — but it is not possible, say, to calculate averages or standard deviations.

Distributions of this kind have to be analysed by means of *non-parametric* methods, which are fewer and less powerful than the interval-scale techniques already mentioned. Chief among them is the Chi-squared test (see Appendix II), which allows us to compare our observations with chance expectation, or with expectations based on specific hypotheses. (Warning: the Chi-squared test must *never* be applied to percentages but only to the raw frequencies in our tables.) There will obviously be problems when we try to correlate two nominal measures with each other, or with a scaled variable, though there are some measures of association that can be used, and there are some non-parametric techniques for the multi-variate analysis of nominal or ordinal data. However, in general the statistical management of nominal measures has fewer and less powerful techniques at its disposal than are available for the management of scaled data.

Between the scaled and the nominal measures lie the *ordinal* or ranked measures. These data do not have interval values but are ranked or ordered on an underlying continuum, such as a pupil's position in class or the results of a race, ranked from first to last. There is a special set of statistical techniques which make use of the ordinal properties of such data. The data do not have interval properties but neither are they completely discrete categories, and this has led to the development of measures of rank correlation, tests of significance for ranked data and so forth. However, these techniques are less powerful and relatively limited in number compared to those available for scaled or metric data.

This may explain why researchers frequently 'bend the rules' in order to be able to use parametric techniques of analysis. Take, for example, a five-point classification of socioeconomic status (SES). Strictly speaking, we have five classes or categories in no particular order, with no numerical scale-interval values, and no underlying scale. It follows that we would have to confine ourselves to the use of Chi-squared for significance testing, and we would not be able to use ordinary (product moment) correlation coefficients to relate SES to scaled data such as income, birthweight or holiday expenditure. However, suppose we made the assumption that socioeconomic status is a linear dimension, from low to high in a straight line; and suppose, further, that we made the assumption that each status category is not independent but has a rank, from first to fifth. In that case, we would have an ordinal scale, which would make our analysis easier and more powerful. Better still, suppose we made the assumption that the status categories are not merely ranked but have interval properties; that would be really helpful. Now we can calculate status averages for different occupational groups, we can use correlations to link SES with numerous scaled values such as educational achievement tests, and we can include SES in major multi-variate analyses to show its importance in 'explaining', say, the outcomes of different forms of psychotherapy.

In strict measurement terms this is wrong, and it follows that the parametric techniques for statistical analysis are not applicable. Yet researchers frequently take liberties with the requirements and assumptions on which these statistical techniques are based in the hope that the techniques will prove sufficiently robust to withstand a certain amount of abuse, and that the ensuing results and conclusions will not be too misleading.

One last word about early thermometry, where practitioners seem to have struggled with difficulties not unlike our own. It is obvious now that the original thermometers were rather like pseudo-scales. The units were arbitrary and had different sizes and different meanings according to different measurement systems; the assumption of the linear expansion rate of mercury and of certain gases was not demonstrated until later; there was either no zero-point or there were arbitrary zero-points located to suit human convenience; and for a long time the readings were unreliable because of technical imperfections in the technique of glass-blowing. Also, it took a great deal of international comparative work and organization before universal standards were fixed for the melting points of various metals or for the average temperature of the healthy human body. New, more accurate and more valid measures only began to develop when scientists acquired a better understanding of the underlying processes that determine temperature change.

The reliability and validity of scaled measures

Up to this point we have been chiefly concerned with two important attributes of scales: *dimensionality* (the presence of a single, linear, underlying continuum) and *equality of intervals* and their statistical treatment. We now turn to a consideration of reliability and validity, two properties which constitute the essence of measurement or data generation of any kind. Each of them is important and, as we have seen previously, they are related to each other.

In the preceding chapter we dealt with reliability and validity in very general terms, and also more specifically in respect of questions dealing with facts and behaviour, and in respect of single questions dealing with attitudes and beliefs. We must now deal more specifically with the reliability and validity of scales and other multi-question approaches.

RELIABILITY

We shall start with reliability, for adequate reliability is a precondition to validity. Reliability means consistency. We need to be sure that the measuring instrument will behave in a fashion which is consistent with itself; that a very high proportion of the score on every occasion is due to the underlying scale variable, with a minimum of error. If we find differences between readings on the same instrument on two separate occasions, or when applied to two different objects or respondents, we must then be sure that these are 'genuine' differences or changes in the subject of measurement, and not differences which could be attributed to inconsistencies in the measuring instrument or to changes in the attendant conditions. The notion of reliability thus includes both the characteristics of the instrument and the conditions under which it is administered — both must be consistent. In terms of Figure 9.1 (page 151), it is the *error component*, the impurity, which produces the inconsistencies and unreliability which we shall seek to reduce. For instance, when comparing lung X-rays of a patient before and after treatment, the doctor must feel confident that the observed differences are real differences in the patient's lungs and not differences due to imperfections in the emulsion, inconsistent exposure times or postural changes of the patient.

Reliability, or self-consistency, is never perfect; it is always a matter of degree. We express it in the form of a correlation coefficient and, in the social and behavioural sciences, it is rare to find reliabilities much above .90. The square of a correlation coefficient expresses the percentage of shared true variance; thus, a reliability coefficient of .90 means that the two measures have .81 or 81 per cent in common — they overlap, or share a common variance, by just over four-fifths. If the reliability of a scale or other measure drops below .80 this means that repeated administrations will cover less than 64 per cent of the same ground, and that the error component is more than one-third; such a measure will come in for serious criticism and might well have to be discarded or rebuilt.

Reliability may be measured in several different ways: (i) by repeatedly administering the scale to the same sample within a short period (*test-retest* reliability). However, this may produce resistance, as well as a practice effect — in a sense, it will no longer mean that the 'same' test is being administered under the 'same' conditions. To avoid these problems we can use (ii) the *internal consistency* method, usually associated with Cronbach's Alpha coefficient and its variants; (iii) the *split-half* method; and (iv) the *parallel-form* method. All these methods yield reliability measures in the form of correlation coefficients. The parallel form method applies to the situation where the test-building procedure has yielded two equivalent forms of the test, identical in every way except for the contents of the items. Having an alternative form of a measuring instrument can be very useful in a before-and-after design, so that respondents will not receive the same measure twice within a short time period. The split-half method meets the latter problem in a different way: the group of items comprising the measure is divided into two halves at random, and the two halves are then intercorrelated.

The internal-consistency method rests firmly on classical scaling theory. If the scale is expected to measure a single underlying continuum, then the items should have strong relationships both with that continuum and with each other. While we cannot observe the former, a scale will be internally consistent if the items correlate highly with each other — in which case they are also more likely to measure the same homogenous variable. Items are more likely to satisfy these requirements if they are reliable, that is if they have low error-components. Since coefficient Alpha gives us an estimate of the proportion of the total variance that is not due to error, this represents the reliability of the scale.

VALIDITY

Validity may also be expressed as a correlation coefficient, but this has a different meaning from that of the reliability coefficient. There are other ways of expressing validity, depending on the type of validation being used. In principle, validity indicates the degree to which an instrument measures what it is supposed or intended to measure, but this is a somewhat vague statement, especially since what we usually seek to measure is an abstraction, so that measurement must take place indirectly. Thus, using our previous example of the doctor looking at two X-ray plates, and assuming that the negatives are reliable and are indeed of the same patient before and after treatment, we note that the doctor is not really interested in the differences between two photographic negatives as such, but in changes in the patient's lungs. If the technique is valid, then there will be close correspondence (concurrent validity) between the X-rays and the changes in the lungs. From these, the doctor will try to draw conclusions about further treatment, prognosis and so forth (predictive validity). If the technique had poor validity, this might lead the doctor to draw erroneous conclusions. Note, however, that at best the technique can only reveal changes in the patient's lungs — the rest is inference. Thus we note that each measurement technique can have more than one validity (indeed, potentially an infinite number of validities), depending on the conclusions we want to draw from it — in this instance, the concurrent validity of changes in the lungs and the predictive validity of estimated recovery rate, prognosis and so on (which may be different). We also note that what we want to measure, and what the instrument is 'supposed to measure', may often be an abstract attribute of some kind (for example 'response to treatment', 'likely outcome') for which, at the time, we may have no other, 'true' criterion measure.

Inevitably, these problems are magnified in the social and behavioural sciences. An achievement test in arithmetic may possibly give us a valid measure of *this* child's ability to solve *this* set of problems *today*. In this limited sense it could be said to be 'valid' because it has concurrent validity, but usually we ask such tests to bear a heavier burden. For example, we may want to treat this set of arithmetical problems as representative of all such problems at this level and to predict that this child could solve *another* set of similar problems tomorrow; that is, we may wish to generalize from this score. How valid is such a generalization likely to be? Or we may wish to use this test to assess the success of a new teaching method which is intended to benefit the less-able pupils; would the test be a valid measure by which to assess the efficacy of the new method? And might it also be able to show which pupils would benefit most: does it have predictive validity?

But how do we know that the test is a test of arithmetic? That it measures 'what it is supposed to measure', an abstraction? A common-sense answer would be to examine the contents of the items. This is known, somewhat disparagingly, as 'face validity' and holds many dangers. For example, out of the entire domain of arithmetic, how have these items been selected? What do they represent? For what level of achievement? How well balanced are the items in terms of content? How pure are the items? In addition to knowing some arithmetic, the child must know how to read and write, how to concentrate and solve problems, how to handle language and so on, and these abilities may, in turn, be related to social background factors. Thus, part of the score will show the child's level of achievement in arithmetic, and another part will measure several other skills, which will not merely introduce some random error into the data but a correlated bias, so children from certain backgrounds will tend to have higher scores. We have to satisfy ourselves that these dangers have somehow been minimized by the way in which the test has been constructed.

One method of establishing concurrent validity might be to correlate the scores with an external criterion — with some other, 'truer' measure — but which one? Teachers' ratings or examination results are often used in such contexts but these tend to be unreliable, and anyway, why are they 'better'? Good criterion measures are notoriously hard to find.

There are other ways of approaching the problem of validity. Sometimes we can use *construct validity*. There are times when we have good theoretical grounds for making certain predictions (for example, that children who score higher on an IQ test will also do better at arithmetic), and so we regard the fulfilment of a set of such predictions as evidence of the test's validity. Certainly we should be rather suspicious of an arithmetic test on which the brighter children did less well (always assuming that the intelligence test was reliable and valid in the first place). Finally, we can examine the test's *predictive validity*; for example, high scorers should be able to give change correctly if they are put behind the counter of a shop, they should get higher marks in an arithmetic examination at the end of the year, and they might become more successful bookmakers or actuaries! An aptitude test for motor mechanics might have high validity for predicting examination results at the end of a training course, but may have poor validity in predicting success as a garage owner, and no-one has yet succeeded in designing a test, or a battery of tests, that would predict success in final university examinations. The usefulness of predictive validity, if attained, can be very great — but we may have to wait a long time to achieve it, and even then we must make sure that the criterion measure is sufficiently valid and reliable for our purposes.

To sum up, a measure may be used for a number of distinct purposes, and it is likely to have a different validity coefficient for each of them. There are also different types of validity:

(i) *content validity*, which seeks to establish that the items or questions are a well-balanced sample of the content domain to be measured ('face validity' is not really good enough);

(ii) *concurrent validity*, which shows how well the test correlates with other, well-validated measures of the same topic, administered at about the same time;

(iii) *predictive validity*, which shows how well the test can forecast some future criterion such as job performance, recovery from illness or future examination attainment;

(iv) *construct validity*, which shows how well the test links up with a set of theoretical assumptions about an abstract construct such as intelligence, conservatism or neuroticism.

As we have seen earlier, validity and reliability are related to each other. Above all, reliability is a necessary (though not sufficient) condition for validity: a measure which is unreliable cannot attain an adequate degree of validity — its error component is too great. On the other hand, a measure may be highly reliable and yet invalid. Consider, for example, three radio direction finders who are homing in on the transmissions from a plane, a ship, or a fox that has been radio-tagged. By triangulating their observations and plotting the results on a map, they will obtain a small triangle (or, in the ideal case, a point) from within which the transmissions must have originated. The smaller the triangle, the more reliable the measurement — but by the time the triangle is drawn, the location of the plane, ship or fox is likely to have changed, perhaps by thirty to fifty miles, and so the measure will no longer be a valid indication of position. We can hope to improve on this situation by taking repeated observations, plotting course and speed and thereby *inferring* a valid location (though with a fox carrying a radio collar, this might not work!). Thus we see that, if we can establish good reliability then we are, as it were, 'in business'; we may not yet have much evidence of validity, but there is hope that with further work the scale's validity may become established at an acceptable level.

In general, the more difficult it becomes to find a suitable external criterion, a strong construct system, or some other method of validation, the more researchers will concentrate on content validity and on making their measures reliable. This is particularly the case in the field of attitude scaling (see below and Chapter 11).

— Stop —

Scores and norms

By itself, a numerical score on a measure will be almost meaningless. To state that someone has scored thirty-two on an ethnic prejudice scale will tell us very little, unless we can place such a score in its *context*. Once a scale has been developed we therefore seek to give it a scoring system and to obtain norms for it. For example, in the case of the test of arithmetical achievement which we mentioned before, we could administer the test to a large number of school children — preferably to representative samples of the relevant age groups — and obtain frequency distributions, percentile curves, means and standard deviations and so on for each school year, and perhaps separately for boys and girls or some other sub-groups of concern. We might also seek to correlate our test with other tests to give more 'meaning' to the scores. Once we have done all this, we can state that a score of, say, forty-nine achieved by a ten-year-old boy is very much below average for children of his age, or that his arithmetic achievement is well below what his IQ test would suggest, so that some remedial work might be needed in order to correct this 'under-achievement'.

In the more complex and multi-faceted domain of attitude measurement we do not generally take such a 'positivist' approach, but it remains necessary to place our scale scores in context, by showing the statistical results of giving the scale to substantial groups (preferably probability samples) of the population of concern.

Improving the reliability of attitude scales

The traditional method of measuring attitudes is by means of *attitude statements*. In the next chapter we shall see the ways in which these may be composed and assembled into an item pool, but for the moment we shall treat them as any other items that form the raw material of scale building. The purpose of such a scale would be to place each respondent on an attitudinal dimension or continuum, but at the start of our research we may not know whether such a dimension exists, whether it is likely to be linear, or whether it might perhaps break up into several different dimensions.

Though in writing the statements or questions we shall obviously be aiming at content validity, we must be acutely aware that each item may contain a substantial error-component which will make it unreliable as a measure, and that there is no way of 'purifying' single items. Thus, suppose we were building some parental attitude scales and that we were working on items for a rejection scale, a way of finding out whether a parent shows a tendency to reject his or her children. This is obviously a subtle problem, but we might try a permissively phrased item such as: 'If we could afford it, we would like to send our children to boarding school'. But how can we know whether this item really belongs in a rejection scale, for instance whether a mother who responds to this item with