

## Measurement and Scaling Techniques

### MEASUREMENT IN RESEARCH

In our daily life we are said to measure when we use some yardstick to determine weight, height, or some other feature of a physical object. We also measure when we judge how well we like a song, a painting or the personalities of our friends. We, thus, measure physical objects as well as abstract concepts. Measurement is a relatively complex and demanding task, specially so when it concerns qualitative or abstract phenomena. By measurement we mean the process of assigning numbers to objects or observations, the level of measurement being a function of the rules under which the numbers are assigned.

It is easy to assign numbers in respect of properties of some objects, but it is relatively difficult in respect of others. For instance, measuring such things as social conformity, intelligence, or marital adjustment is much less obvious and requires much closer attention than measuring physical weight, biological age or a person's financial assets. In other words, properties like weight, height, etc., can be measured directly with some standard unit of measurement, but it is not that easy to measure properties like motivation to succeed, ability to stand stress and the like. We can expect high accuracy in measuring the length of pipe with a yard stick, but if the concept is abstract and the measurement tools are not standardized, we are less confident about the accuracy of the results of measurement.

Technically speaking, measurement is a process of mapping aspects of a domain onto other aspects of a range according to some rule of correspondence. In measuring, we devise some form of scale in the range (in terms of set theory, range may refer to some set) and then transform or map the properties of objects from the domain (in terms of set theory, domain may refer to some other set) onto this scale. For example, in case we are to find the male to female attendance ratio while conducting a study of persons who attend some show, then we may tabulate those who come to the show according to sex. In terms of set theory, this process is one of mapping the observed physical properties of those coming to the show (the domain) on to a sex classification (the range). The rule of correspondence is: If the object in the domain appears to be male, assign to "0" and if female assign to "1". Similarly, we can record a person's marital status as 1, 2, 3 or 4, depending on whether

the person is single, married, widowed or divorced. We can as well record “Yes or No” answers to a question as “0” and “1” (or as 1 and 2 or perhaps as 59 and 60). In this artificial or nominal way, categorical data (qualitative or descriptive) can be made into numerical data and if we thus code the various categories, we refer to the numbers we record as nominal data. *Nominal data* are numerical in name only, because they do not share any of the properties of the numbers we deal in ordinary arithmetic. For instance if we record marital status as 1, 2, 3, or 4 as stated above, we cannot write  $4 > 2$  or  $3 < 4$  and we cannot write  $3 - 1 = 4 - 2$ ,  $1 + 3 = 4$  or  $4 \div 2 = 2$ .

In those situations when we cannot do anything except set up inequalities, we refer to the data as *ordinal data*. For instance, if one mineral can scratch another, it receives a higher hardness number and on Mohs’ scale the numbers from 1 to 10 are assigned respectively to talc, gypsum, calcite, fluorite, apatite, feldspar, quartz, topaz, sapphire and diamond. With these numbers we can write  $5 > 2$  or  $6 < 9$  as apatite is harder than gypsum and feldspar is softer than sapphire, but we cannot write for example  $10 - 9 = 5 - 4$ , because the difference in hardness between diamond and sapphire is actually much greater than that between apatite and fluorite. It would also be meaningless to say that topaz is twice as hard as fluorite simply because their respective hardness numbers on Mohs’ scale are 8 and 4. The greater than symbol (i.e.,  $>$ ) in connection with ordinal data may be used to designate “happier than” “preferred to” and so on.

When in addition to setting up inequalities we can also form differences, we refer to the data as *interval data*. Suppose we are given the following temperature readings (in degrees Fahrenheit):  $58^\circ$ ,  $63^\circ$ ,  $70^\circ$ ,  $95^\circ$ ,  $110^\circ$ ,  $126^\circ$  and  $135^\circ$ . In this case, we can write  $100^\circ > 70^\circ$  or  $95^\circ < 135^\circ$  which simply means that  $110^\circ$  is warmer than  $70^\circ$  and that  $95^\circ$  is cooler than  $135^\circ$ . We can also write for example  $95^\circ - 70^\circ = 135^\circ - 110^\circ$ , since equal temperature differences are equal in the sense that the same amount of heat is required to raise the temperature of an object from  $70^\circ$  to  $95^\circ$  or from  $110^\circ$  to  $135^\circ$ . On the other hand, it would not mean much if we said that  $126^\circ$  is twice as hot as  $63^\circ$ , even though  $126^\circ \div 63^\circ = 2$ . To show the reason, we have only to change to the centigrade scale, where the first temperature becomes  $5/9 (126 - 32) = 52^\circ$ , the second temperature becomes  $5/9 (63 - 32) = 17^\circ$  and the first figure is now more than three times the second. This difficulty arises from the fact that Fahrenheit and Centigrade scales both have artificial origins (zeros) i.e., the number 0 of neither scale is indicative of the absence of whatever quantity we are trying to measure.

When in addition to setting up inequalities and forming differences we can also form quotients (i.e., when we can perform all the customary operations of mathematics), we refer to such data as *ratio data*. In this sense, ratio data includes all the usual measurement (or determinations) of length, height, money amounts, weight, volume, area, pressures etc.

The above stated distinction between nominal, ordinal, interval and ratio data is important for the nature of a set of data may suggest the use of particular statistical techniques\*. A researcher has to be quite alert about this aspect while measuring properties of objects or of abstract concepts.

\* When data can be measured in units which are interchangeable e.g., weights (by ratio scales), temperatures (by interval scales), that data is said to be parametric and can be subjected to most kinds of statistical and mathematical processes. But when data is measured in units which are not interchangeable, e.g., product preferences (by ordinal scales), the data is said to be non-parametric and is susceptible only to a limited extent to mathematical and statistical treatment.

## MEASUREMENT SCALES

From what has been stated above, we can write that scales of measurement can be considered in terms of their mathematical properties. The most widely used classification of measurement scales are: (a) nominal scale; (b) ordinal scale; (c) interval scale; and (d) ratio scale.

**(a) Nominal scale:** Nominal scale is simply a system of assigning number symbols to events in order to label them. The usual example of this is the assignment of numbers of basketball players in order to identify them. Such numbers cannot be considered to be associated with an ordered scale for their order is of no consequence; the numbers are just convenient labels for the particular class of events and as such have no quantitative value. Nominal scales provide convenient ways of keeping track of people, objects and events. One cannot do much with the numbers involved. For example, one cannot usefully average the numbers on the back of a group of football players and come up with a meaningful value. Neither can one usefully compare the numbers assigned to one group with the numbers assigned to another. The counting of members in each group is the only possible arithmetic operation when a nominal scale is employed. Accordingly, we are restricted to use mode as the measure of central tendency. There is no generally used measure of dispersion for nominal scales. Chi-square test is the most common test of statistical significance that can be utilized, and for the measures of correlation, the contingency coefficient can be worked out.

Nominal scale is the least powerful level of measurement. It indicates no order or distance relationship and has no arithmetic origin. A nominal scale simply describes differences between things by assigning them to categories. Nominal data are, thus, counted data. The scale wastes any information that we may have about varying degrees of attitude, skills, understandings, etc. In spite of all this, nominal scales are still very useful and are widely used in surveys and other *ex-post-facto* research when data are being classified by major sub-groups of the population.

**(b) Ordinal scale:** The lowest level of the ordered scale that is commonly used is the ordinal scale. The ordinal scale places events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule. Rank orders represent ordinal scales and are frequently used in research relating to qualitative phenomena. A student's rank in his graduation class involves the use of an ordinal scale. One has to be very careful in making statement about scores based on ordinal scales. For instance, if Ram's position in his class is 10 and Mohan's position is 40, it cannot be said that Ram's position is four times as good as that of Mohan. The statement would make no sense at all. Ordinal scales only permit the ranking of items from highest to lowest. Ordinal measures have no absolute values, and the real differences between adjacent ranks may not be equal. All that can be said is that one person is higher or lower on the scale than another, but more precise comparisons cannot be made.

Thus, the use of an ordinal scale implies a statement of 'greater than' or 'less than' (an equality statement is also acceptable) without our being able to state how much greater or less. The real difference between ranks 1 and 2 may be more or less than the difference between ranks 5 and 6. Since the numbers of this scale have only a rank meaning, the appropriate measure of central tendency is the median. A percentile or quartile measure is used for measuring dispersion. Correlations are restricted to various rank order methods. Measures of statistical significance are restricted to the non-parametric methods.

**(c) Interval scale:** In the case of interval scale, the intervals are adjusted in terms of some rule that has been established as a basis for making the units equal. The units are equal only in so far as one

accepts the assumptions on which the rule is based. Interval scales can have an arbitrary zero, but it is not possible to determine for them what may be called an absolute zero or the unique origin. The primary limitation of the interval scale is the lack of a true zero; it does not have the capacity to measure the complete absence of a trait or characteristic. The Fahrenheit scale is an example of an interval scale and shows similarities in what one can and cannot do with it. One can say that an increase in temperature from 30° to 40° involves the same increase in temperature as an increase from 60° to 70°, but one cannot say that the temperature of 60° is twice as warm as the temperature of 30° because both numbers are dependent on the fact that the zero on the scale is set arbitrarily at the temperature of the freezing point of water. The ratio of the two temperatures, 30° and 60°, means nothing because zero is an arbitrary point.

Interval scales provide more powerful measurement than ordinal scales for interval scale also incorporates the concept of equality of interval. As such more powerful statistical measures can be used with interval scales. Mean is the appropriate measure of central tendency, while standard deviation is the most widely used measure of dispersion. Product moment correlation techniques are appropriate and the generally used tests for statistical significance are the 't' test and 'F' test.

**(d) Ratio scale:** Ratio scales have an absolute or true zero of measurement. The term 'absolute zero' is not as precise as it was once believed to be. We can conceive of an absolute zero of length and similarly we can conceive of an absolute zero of time. For example, the zero point on a centimeter scale indicates the complete absence of length or height. But an absolute zero of temperature is theoretically unobtainable and it remains a concept existing only in the scientist's mind. The number of minor traffic-rule violations and the number of incorrect letters in a page of type script represent scores on ratio scales. Both these scales have absolute zeros and as such all minor traffic violations and all typing errors can be assumed to be equal in significance. With ratio scales involved one can make statements like "Jyoti's" typing performance was twice as good as that of "Reetu." The ratio involved does have significance and facilitates a kind of comparison which is not possible in case of an interval scale.

Ratio scale represents the actual amounts of variables. Measures of physical dimensions such as weight, height, distance, etc. are examples. Generally, all statistical techniques are usable with ratio scales and all manipulations that one can carry out with real numbers can also be carried out with ratio scale values. Multiplication and division can be used with this scale but not with other scales mentioned above. Geometric and harmonic means can be used as measures of central tendency and coefficients of variation may also be calculated.

Thus, proceeding from the nominal scale (the least precise type of scale) to ratio scale (the most precise), relevant information is obtained increasingly. If the nature of the variables permits, the researcher should use the scale that provides the most precise description. Researchers in physical sciences have the advantage to describe variables in ratio scale form but the behavioural sciences are generally limited to describe variables in interval scale form, a less precise type of measurement.

### Sources of Error in Measurement

Measurement should be precise and unambiguous in an ideal research study. This objective, however, is often not met with in entirety. As such the researcher must be aware about the sources of error in measurement. The following are the possible sources of error in measurement.

**(a) Respondent:** At times the respondent may be reluctant to express strong negative feelings or it is just possible that he may have very little knowledge but may not admit his ignorance. All this reluctance is likely to result in an interview of 'guesses.' Transient factors like fatigue, boredom, anxiety, etc. may limit the ability of the respondent to respond accurately and fully.

**(b) Situation:** Situational factors may also come in the way of correct measurement. Any condition which places a strain on interview can have serious effects on the interviewer-respondent rapport. For instance, if someone else is present, he can distort responses by joining in or merely by being present. If the respondent feels that anonymity is not assured, he may be reluctant to express certain feelings.

**(c) Measurer:** The interviewer can distort responses by rewording or reordering questions. His behaviour, style and looks may encourage or discourage certain replies from respondents. Careless mechanical processing may distort the findings. Errors may also creep in because of incorrect coding, faulty tabulation and/or statistical calculations, particularly in the data-analysis stage.

**(d) Instrument:** Error may arise because of the defective measuring instrument. The use of complex words, beyond the comprehension of the respondent, ambiguous meanings, poor printing, inadequate space for replies, response choice omissions, etc. are a few things that make the measuring instrument defective and may result in measurement errors. Another type of instrument deficiency is the poor sampling of the universe of items of concern.

Researcher must know that correct measurement depends on successfully meeting all of the problems listed above. He must, to the extent possible, try to eliminate, neutralize or otherwise deal with all the possible sources of error so that the final results may not be contaminated.

## Tests of Sound Measurement

Sound measurement must meet the tests of validity, reliability and practicality. In fact, these are the three major considerations one should use in evaluating a measurement tool. "Validity refers to the extent to which a test measures what we actually wish to measure. Reliability has to do with the accuracy and precision of a measurement procedure ... Practicality is concerned with a wide range of factors of economy, convenience, and interpretability ..." <sup>1</sup> We briefly take up the relevant details concerning these tests of sound measurement.

### 1. Test of Validity\*

Validity is the most critical criterion and indicates the degree to which an instrument measures what it is supposed to measure. Validity can also be thought of as utility. In other words, validity is the extent to which differences found with a measuring instrument reflect true differences among those being tested. But the question arises: how can one determine validity without direct confirming knowledge? The answer may be that we seek other relevant evidence that confirms the answers we have found with our measuring tool. What is relevant, evidence often depends upon the nature of the

<sup>1</sup> Robert L. Thorndike and Elizabeth Hagen: *Measurement and Evaluation in Psychology and Education*, 3rd Ed., p. 162.

\* Two forms of validity are usually mentioned in research literature viz., the external validity and the internal validity. External validity of research findings is their generalizability to populations, settings, treatment variables and measurement variables. We shall talk about it in the context of significance tests later on. The internal validity of a research design is its ability to measure what it aims to measure. We shall deal with this validity only in the present chapter.

research problem and the judgement of the researcher. But one can certainly consider three types of validity in this connection: (i) Content validity; (ii) Criterion-related validity and (iii) Construct validity.

(i) *Content validity* is the extent to which a measuring instrument provides adequate coverage of the topic under study. If the instrument contains a representative sample of the universe, the content validity is good. Its determination is primarily judgemental and intuitive. It can also be determined by using a panel of persons who shall judge how well the measuring instrument meets the standards, but there is no numerical way to express it.

(ii) *Criterion-related validity* relates to our ability to predict some outcome or estimate the existence of some current condition. This form of validity reflects the success of measures used for some empirical estimating purpose. The concerned criterion must possess the following qualities:

*Relevance:* (A criterion is relevant if it is defined in terms we judge to be the proper measure.)

*Freedom from bias:* (Freedom from bias is attained when the criterion gives each subject an equal opportunity to score well.)

*Reliability:* (A reliable criterion is stable or reproducible.)

*Availability:* (The information specified by the criterion must be available.)

In fact, a Criterion-related validity is a broad term that actually refers to (i) *Predictive validity* and (ii) *Concurrent validity*. The former refers to the usefulness of a test in predicting some future performance whereas the latter refers to the usefulness of a test in closely relating to other measures of known validity. Criterion-related validity is expressed as the coefficient of correlation between test scores and some measure of future performance or between test scores and scores on another measure of known validity.

(iii) *Construct validity* is the most complex and abstract. A measure is said to possess construct validity to the degree that it confirms to predicted correlations with other theoretical propositions. Construct validity is the degree to which scores on a test can be accounted for by the explanatory constructs of a sound theory. For determining construct validity, we associate a set of other propositions with the results received from using our measurement instrument. If measurements on our devised scale correlate in a predicted way with these other propositions, we can conclude that there is some construct validity.

If the above stated criteria and tests are met with, we may state that our measuring instrument is valid and will result in correct measurement; otherwise we shall have to look for more information and/or resort to exercise of judgement.

## 2. Test of Reliability

The test of reliability is another important test of sound measurement. A measuring instrument is reliable if it provides consistent results. Reliable measuring instrument does contribute to validity, but a reliable instrument need not be a valid instrument. For instance, a scale that consistently overweighs objects by five kgs., is a reliable scale, but it does not give a valid measure of weight. But the other way is not true i.e., a valid instrument is always reliable. Accordingly reliability is not as valuable as validity, but it is easier to assess reliability in comparison to validity. If the quality of reliability is satisfied by an instrument, then while using it we can be confident that the transient and situational factors are not interfering.



Two aspects of reliability viz., stability and equivalence deserve special mention. The *stability aspect* is concerned with securing consistent results with repeated measurements of the same person and with the same instrument. We usually determine the degree of stability by comparing the results of repeated measurements. The *equivalence aspect* considers how much error may get introduced by different investigators or different samples of the items being studied. A good way to test for the equivalence of measurements by two investigators is to compare their observations of the same events. Reliability can be improved in the following two ways:

- (i) By standardising the conditions under which the measurement takes place i.e., we must ensure that external sources of variation such as boredom, fatigue, etc., are minimised to the extent possible. That will improve stability aspect.
- (ii) By carefully designed directions for measurement with no variation from group to group, by using trained and motivated persons to conduct the research and also by broadening the sample of items used. This will improve equivalence aspect.

### 3. Test of Practicality

The practicality characteristic of a measuring instrument can be judged in terms of economy, convenience and interpretability. From the operational point of view, the measuring instrument ought to be practical i.e., it should be economical, convenient and interpretable. *Economy* consideration suggests that some trade-off is needed between the ideal research project and that which the budget can afford. The length of measuring instrument is an important area where economic pressures are quickly felt. Although more items give greater reliability as stated earlier, but in the interest of limiting the interview or observation time, we have to take only few items for our study purpose. Similarly, data-collection methods to be used are also dependent at times upon economic factors. *Convenience* test suggests that the measuring instrument should be easy to administer. For this purpose one should give due attention to the proper layout of the measuring instrument. For instance, a questionnaire, with clear instructions (illustrated by examples), is certainly more effective and easier to complete than one which lacks these features. *Interpretability* consideration is specially important when persons other than the designers of the test are to interpret the results. The measuring instrument, in order to be interpretable, must be supplemented by (a) detailed instructions for administering the test; (b) scoring keys; (c) evidence about the reliability and (d) guides for using the test and for interpreting results.

## TECHNIQUE OF DEVELOPING MEASUREMENT TOOLS

The technique of developing measurement tools involves a four-stage process, consisting of the following:

- (a) Concept development;
- (b) Specification of concept dimensions;
- (c) Selection of indicators; and
- (d) Formation of index.

The first and foremost step is that of *concept development* which means that the researcher should arrive at an understanding of the major concepts pertaining to his study. This step of concept

development is more apparent in theoretical studies than in the more pragmatic research, where the fundamental concepts are often already established.

The second step requires the researcher to specify the *dimensions of the concepts* that he developed in the first stage. This task may either be accomplished by deduction i.e., by adopting a more or less intuitive approach or by empirical correlation of the individual dimensions with the total concept and/or the other concepts. For instance, one may think of several dimensions such as product reputation, customer treatment, corporate leadership, concern for individuals, sense of social responsibility and so forth when one is thinking about the image of a certain company.

Once the dimensions of a concept have been specified, the researcher must *develop indicators* for measuring each concept element. Indicators are specific questions, scales, or other devices by which respondent's knowledge, opinion, expectation, etc., are measured. As there is seldom a perfect measure of a concept, the researcher should consider several alternatives for the purpose. The use of more than one indicator gives stability to the scores and it also improves their validity.

The last step is that of combining the various indicators into an index, i.e., *formation of an index*. When we have several dimensions of a concept or different measurements of a dimension, we may need to combine them into a single index. One simple way for getting an overall index is to provide scale values to the responses and then sum up the corresponding scores. Such an overall index would provide a better measurement tool than a single indicator because of the fact that an "individual indicator has only a probability relation to what we really want to know."<sup>2</sup> This way we must obtain an overall index for the various concepts concerning the research study.

## Scaling

In research we quite often face measurement problem (since we want a valid measurement but may not obtain it), specially when the concepts to be measured are complex and abstract and we do not possess the standardised measurement tools. Alternatively, we can say that while measuring attitudes and opinions, we face the problem of their valid measurement. Similar problem may be faced by a researcher, of course in a lesser degree, while measuring physical or institutional concepts. As such we should study some procedures which may enable us to measure abstract concepts more accurately. This brings us to the study of scaling techniques.

## Meaning of Scaling

Scaling describes the procedures of assigning numbers to various degrees of opinion, attitude and other concepts. This can be done in two ways viz., (i) making a judgement about some characteristic of an individual and then placing him directly on a scale that has been defined in terms of that characteristic and (ii) constructing questionnaires in such a way that the score of individual's responses assigns him a place on a scale. It may be stated here that a scale is a continuum, consisting of the highest point (in terms of some characteristic e.g., preference, favourableness, etc.) and the lowest point along with several intermediate points between these two extreme points. These scale-point positions are so related to each other that when the first point happens to be the highest point, the second point indicates a higher degree in terms of a given characteristic as compared to the third

<sup>2</sup> Lazarsfeld, *Evidence and Inference*, p. 112.



point and the third point indicates a higher degree as compared to the fourth and so on. Numbers for measuring the distinctions of degree in the attitudes/opinions are, thus, assigned to individuals corresponding to their scale-positions. All this is better understood when we talk about scaling technique(s). Hence the term 'scaling' is applied to the procedures for attempting to determine quantitative measures of subjective abstract concepts. Scaling has been defined as a "procedure for the assignment of numbers (or other symbols) to a property of objects in order to impart some of the characteristics of numbers to the properties in question."<sup>3</sup>

### Scale Classification Bases

The number assigning procedures or the scaling procedures may be broadly classified on one or more of the following bases: (a) subject orientation; (b) response form; (c) degree of subjectivity; (d) scale properties; (e) number of dimensions and (f) scale construction techniques. We take up each of these separately.

**(a) Subject orientation:** Under it a scale may be designed to measure characteristics of the respondent who completes it or to judge the stimulus object which is presented to the respondent. In respect of the former, we presume that the stimuli presented are sufficiently homogeneous so that the between-stimuli variation is small as compared to the variation among respondents. In the latter approach, we ask the respondent to judge some specific object in terms of one or more dimensions and we presume that the between-respondent variation will be small as compared to the variation among the different stimuli presented to respondents for judging.

**(b) Response form:** Under this we may classify the scales as categorical and comparative. Categorical scales are also known as rating scales. These scales are used when a respondent scores some object without direct reference to other objects. Under comparative scales, which are also known as ranking scales, the respondent is asked to compare two or more objects. In this sense the respondent may state that one object is superior to the other or that three models of pen rank in order 1, 2 and 3. The essence of ranking is, in fact, a relative comparison of a certain property of two or more objects.

**(c) Degree of subjectivity:** With this basis the scale data may be based on whether we measure subjective personal preferences or simply make non-preference judgements. In the former case, the respondent is asked to choose which person he favours or which solution he would like to see employed, whereas in the latter case he is simply asked to judge which person is more effective in some aspect or which solution will take fewer resources without reflecting any personal preference.

**(d) Scale properties:** Considering scale properties, one may classify the scales as nominal, ordinal, interval and ratio scales. Nominal scales merely classify without indicating order, distance or unique origin. Ordinal scales indicate magnitude relationships of 'more than' or 'less than', but indicate no distance or unique origin. Interval scales have both order and distance values, but no unique origin. Ratio scales possess all these features.

**(e) Number of dimensions:** In respect of this basis, scales can be classified as 'unidimensional' and 'multidimensional' scales. Under the former we measure only one attribute of the respondent or object, whereas multidimensional scaling recognizes that an object might be described better by using the concept of an attribute space of 'n' dimensions, rather than a single-dimension continuum.

<sup>3</sup> Bernard S. Phillips, *Social Research Strategy and Tactics*, 2nd ed., p. 205.

**(f) Scale construction techniques:** Following are the five main techniques by which scales can be developed.

- (i) *Arbitrary approach:* It is an approach where scale is developed on *ad hoc* basis. This is the most widely used approach. It is presumed that such scales measure the concepts for which they have been designed, although there is little evidence to support such an assumption.
- (ii) *Consensus approach:* Here a panel of judges evaluate the items chosen for inclusion in the instrument in terms of whether they are relevant to the topic area and unambiguous in implication.
- (iii) *Item analysis approach:* Under it a number of individual items are developed into a test which is given to a group of respondents. After administering the test, the total scores are calculated for every one. Individual items are then analysed to determine which items discriminate between persons or objects with high total scores and those with low scores.
- (iv) *Cumulative scales* are chosen on the basis of their conforming to some ranking of items with ascending and descending discriminating power. For instance, in such a scale the endorsement of an item representing an extreme position should also result in the endorsement of all items indicating a less extreme position.
- (v) *Factor scales* may be constructed on the basis of intercorrelations of items which indicate that a common factor accounts for the relationship between items. This relationship is typically measured through factor analysis method.

### Important Scaling Techniques

We now take up some of the important scaling techniques often used in the context of research specially in context of social or business research.

**Rating scales:** The rating scale involves qualitative description of a limited number of aspects of a thing or of traits of a person. When we use rating scales (or categorical scales), we judge an object in absolute terms against some specified criteria i.e., we judge properties of objects without reference to other similar objects. These ratings may be in such forms as “like-dislike”, “above average, average, below average”, or other classifications with more categories such as “like very much—like some what—neutral—dislike somewhat—dislike very much”; “excellent—good—average—below average—poor”, “always—often—occasionally—rarely—never”, and so on. There is no specific rule whether to use a two-points scale, three-points scale or scale with still more points. In practice, three to seven points scales are generally used for the simple reason that more points on a scale provide an opportunity for greater sensitivity of measurement.

Rating scale may be either a graphic rating scale or an itemized rating scale.

- (i) *The graphic rating scale* is quite simple and is commonly used in practice. Under it the various points are usually put along the line to form a continuum and the rater indicates his rating by simply making a mark (such as ✓) at the appropriate point on a line that runs from one extreme to the other. Scale-points with brief descriptions may be indicated along the line, their function being to assist the rater in performing his job. The following is an example of five-points graphic rating scale when we wish to ascertain people’s liking or disliking any product:

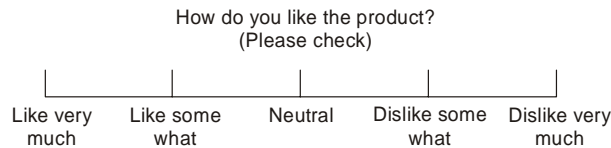


Fig. 5.1

This type of scale has several limitations. The respondents may check at almost any position along the line which fact may increase the difficulty of analysis. The meanings of the terms like “very much” and “some what” may depend upon respondent’s frame of reference so much so that the statement might be challenged in terms of its equivalency. Several other rating scale variants (e.g., boxes replacing line) may also be used.

- (ii) The *itemized rating scale* (also known as numerical scale) presents a series of statements from which a respondent selects one as best reflecting his evaluation. These statements are ordered progressively in terms of more or less of some property. An example of itemized scale can be given to illustrate it.

Suppose we wish to inquire as to how well does a worker get along with his fellow workers? In such a situation we may ask the respondent to select one, to express his opinion, from the following:

- He is almost always involved in some friction with a fellow worker.
- He is often at odds with one or more of his fellow workers.
- He sometimes gets involved in friction.
- He infrequently becomes involved in friction with others.
- He almost never gets involved in friction with fellow workers.

The chief merit of this type of scale is that it provides more information and meaning to the rater, and thereby increases reliability. This form is relatively difficult to develop and the statements may not say exactly what the respondent would like to express.

Rating scales have certain good points. The results obtained from their use compare favourably with alternative methods. They require less time, are interesting to use and have a wide range of applications. Besides, they may also be used with a large number of properties or variables. But their value for measurement purposes depends upon the assumption that the respondents can and do make good judgements. If the respondents are not very careful while rating, errors may occur. Three types of errors are common viz., the error of leniency, the error of central tendency and the error of halo effect. The error of leniency occurs when certain respondents are either easy raters or hard raters. When raters are reluctant to give extreme judgements, the result is the error of central tendency. The error of halo effect or the systematic bias occurs when the rater carries over a generalised impression of the subject from one rating to another. This sort of error takes place when we conclude for example, that a particular report is good because we like its form or that someone is intelligent because he agrees with us or has a pleasing personality. In other words, halo effect is likely to appear when the rater is asked to rate many factors, on a number of which he has no evidence for judgement.

**Ranking scales:** Under ranking scales (or comparative scales) we make relative judgements against other similar objects. The respondents under this method directly compare two or more objects and make choices among them. There are two generally used approaches of ranking scales viz.

**(a) Method of paired comparisons:** Under it the respondent can express his attitude by making a choice between two objects, say between a new flavour of soft drink and an established brand of drink. But when there are more than two stimuli to judge, the number of judgements required in a paired comparison is given by the formula:

$$N = \frac{n(n-1)}{2}$$

where  $N$  = number of judgements

$n$  = number of stimuli or objects to be judged.

For instance, if there are ten suggestions for bargaining proposals available to a workers union, there are 45 paired comparisons that can be made with them. When  $N$  happens to be a big figure, there is the risk of respondents giving ill considered answers or they may even refuse to answer. We can reduce the number of comparisons per respondent either by presenting to each one of them only a sample of stimuli or by choosing a few objects which cover the range of attractiveness at about equal intervals and then comparing all other stimuli to these few standard objects. Thus, paired-comparison data may be treated in several ways. If there is substantial consistency, we will find that if  $X$  is preferred to  $Y$ , and  $Y$  to  $Z$ , then  $X$  will consistently be preferred to  $Z$ . If this is true, we may take the total number of preferences among the comparisons as the score for that stimulus.

It should be remembered that paired comparison provides ordinal data, but the same may be converted into an interval scale by the method of the *Law of Comparative Judgement* developed by L.L. Thurstone. This technique involves the conversion of frequencies of preferences into a table of proportions which are then transformed into  $Z$  matrix by referring to the table of area under the normal curve. J.P. Guilford in his book “Psychometric Methods” has given a procedure which is relatively easier. The method is known as the *Composite Standard Method* and can be illustrated as under:

Suppose there are four proposals which some union bargaining committee is considering. The committee wants to know how the union membership ranks these proposals. For this purpose a sample of 100 members might express the views as shown in the following table:

**Table 5.1:** Response Patterns of 100 Members’ Paired Comparisons of 4 Suggestions for Union Bargaining Proposal Priorities

|        | Suggestion |     |     |     |
|--------|------------|-----|-----|-----|
|        | A          | B   | C   | D   |
| A      | –          | 65* | 32  | 20  |
| B      | 40         | –   | 38  | 42  |
| C      | 45         | 50  | –   | 70  |
| D      | 80         | 20  | 98  | –   |
| TOTAL: | 165        | 135 | 168 | 132 |

\*Read as 65 members preferred suggestion B to suggestion A.

Contd.

| Rank order | 2      | 3      | 1      | 4      |
|------------|--------|--------|--------|--------|
| $M_p$      | 0.5375 | 0.4625 | 0.5450 | 0.4550 |
| $Z_j$      | 0.09   | (-).09 | 0.11   | (-).11 |
| $R_j$      | 0.20   | 0.02   | 0.22   | 0.00   |

Comparing the total number of preferences for each of the four proposals, we find that *C* is the most popular, followed by *A*, *B* and *D* respectively in popularity. The rank order shown in the above table explains all this.

By following the composite standard method, we can develop an interval scale from the paired-comparison ordinal data given in the above table for which purpose we have to adopt the following steps in order:

- (i) Using the data in the above table, we work out the column mean with the help of the formula given below:

$$M_p = \frac{C + .5(N)}{nN} = \frac{165 + .5(100)}{4(100)} = .5375$$

where

$M_p$  = the mean proportion of the columns

$C$  = the total number of choices for a given suggestion

$n$  = number of stimuli (proposals in the given problem)

$N$  = number of items in the sample.

The column means have been shown in the  $M_p$  row in the above table.

- (ii) The  $Z$  values for the  $M_p$  are secured from the table giving the area under the normal curve. When the  $M_p$  value is less than .5, the  $Z$  value is negative and for all  $M_p$  values higher than .5, the  $Z$  values are positive.\* These  $Z$  values are shown in  $Z_j$  row in the above table.
- (iii) As the  $Z_j$  values represent an interval scale, zero is an arbitrary value. Hence we can eliminate negative scale values by giving the value of zero to the lowest scale value (this being (-).11 in our example which we shall take equal to zero) and then adding the absolute value of this lowest scale value to all other scale items. This scale has been shown in  $R_j$  row in the above table.

Graphically we can show this interval scale that we have derived from the paired-comparison data using the composite standard method as follows:

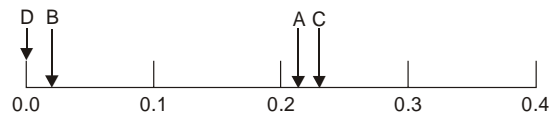


Fig. 5.2

\*To use Normal curve area table for this sort of transformation, we must subtract 0.5 from all  $M_p$  values which exceed .5 to secure the values with which to enter the normal curve area table for which  $Z$  values can be obtained. For all  $M_p$  values of less than .5 we must subtract all such values from 0.5 to secure the values with which to enter the normal curve area table for which  $Z$  values can be obtained but the  $Z$  values in this situation will be with negative sign.

**(b) Method of rank order:** Under this method of comparative scaling, the respondents are asked to rank their choices. This method is easier and faster than the method of paired comparisons stated above. For example, with 10 items it takes 45 pair comparisons to complete the task, whereas the method of rank order simply requires ranking of 10 items only. The problem of transitivity (such as *A* prefers to *B*, *B* to *C*, but *C* prefers to *A*) is also not there in case we adopt method of rank order. Moreover, a complete ranking at times is not needed in which case the respondents may be asked to rank only their first, say, four choices while the number of overall items involved may be more than four, say, it may be 15 or 20 or more. To secure a simple ranking of all items involved we simply total rank values received by each item. There are methods through which we can as well develop an interval scale of these data. But then there are limitations of this method. The first one is that data obtained through this method are ordinal data and hence rank ordering is an ordinal scale with all its limitations. Then there may be the problem of respondents becoming careless in assigning ranks particularly when there are many (usually more than 10) items.

### Scale Construction Techniques

In social science studies, while measuring attitudes of the people we generally follow the technique of preparing the opinionnaire\* (or attitude scale) in such a way that the score of the individual responses assigns him a place on a scale. Under this approach, the respondent expresses his agreement or disagreement with a number of statements relevant to the issue. While developing such statements, the researcher must note the following two points:

- (i) That the statements must elicit responses which are psychologically related to the attitude being measured;
- (ii) That the statements need be such that they discriminate not merely between extremes of attitude but also among individuals who differ slightly.

Researchers must as well be aware that inferring attitude from what has been recorded in opinionnaires has several limitations. People may conceal their attitudes and express socially acceptable opinions. They may not really know how they feel about a social issue. People may be unaware of their attitude about an abstract situation; until confronted with a real situation, they may be unable to predict their reaction. Even behaviour itself is at times not a true indication of attitude. For instance, when politicians kiss babies, their behaviour may not be a true expression of affection toward infants. Thus, there is no sure method of measuring attitude; we only try to measure the expressed opinion and then draw inferences from it about people's real feelings or attitudes.

With all these limitations in mind, psychologists and sociologists have developed several scale construction techniques for the purpose. The researcher should know these techniques so as to develop an appropriate scale for his own study. Some of the important approaches, along with the corresponding scales developed under each approach to measure attitude are as follows:

\* An information form that attempts to measure the attitude or belief of an individual is known as opinionnaire.



**Table 5.2:** Different Scales for Measuring Attitudes of People

| <i>Name of the scale construction approach</i> | <i>Name of the scale developed</i>  |
|--|---|
| 1. Arbitrary approach                          | Arbitrary scales  |
| 2. Consensus scale approach                    | Differential scales (such as Thurstone Differential scale)                              |
| 3. Item analysis approach                      | Summated scales (such as Likert Scale)  |
| 4. Cumulative scale approach                   | Cumulative scales (such as Guttman's Scalogram)   |
| 5. Factor analysis approach                    | Factor scales (such as Osgood's Semantic Differential, Multi-dimensional Scaling, etc.) |

A brief description of each of the above listed scales will be helpful.

### **Arbitrary Scales**

Arbitrary scales are developed on *ad hoc* basis and are designed largely through the researcher's own subjective selection of items. The researcher first collects few statements or items which he believes are unambiguous and appropriate to a given topic. Some of these are selected for inclusion in the measuring instrument and then people are asked to check in a list the statements with which they agree.

The chief merit of such scales is that they can be developed very easily, quickly and with relatively less expense. They can also be designed to be highly specific and adequate. Because of these benefits, such scales are widely used in practice.

At the same time there are some limitations of these scales. The most important one is that we do not have objective evidence that such scales measure the concepts for which they have been developed. We have simply to rely on researcher's insight and competence.

### **Differential Scales (or Thurstone-type Scales)**

The name of L.L. Thurstone is associated with differential scales which have been developed using consensus scale approach. Under such an approach the selection of items is made by a panel of judges who evaluate the items in terms of whether they are relevant to the topic area and unambiguous in implication. The detailed procedure is as under:

- (a) The researcher gathers a large number of statements, usually twenty or more, that express various points of view toward a group, institution, idea, or practice (i.e., statements belonging to the topic area).
- (b) These statements are then submitted to a panel of judges, each of whom arranges them in eleven groups or piles ranging from one extreme to another in position. Each of the judges is requested to place generally in the first pile the statements which he thinks are most unfavourable to the issue, in the second pile to place those statements which he thinks are next most unfavourable and he goes on doing so in this manner till in the eleventh pile he puts the statements which he considers to be the most favourable.
- (c) This sorting by each judge yields a composite position for each of the items. In case of marked disagreement between the judges in assigning a position to an item, that item is discarded.

- (d) For items that are retained, each is given its median scale value between one and eleven as established by the panel. In other words, the scale value of any one statement is computed as the 'median' position to which it is assigned by the group of judges.
- (e) A final selection of statements is then made. For this purpose a sample of statements, whose median scores are spread evenly from one extreme to the other is taken. The statements so selected, constitute the final scale to be administered to respondents. The position of each statement on the scale is the same as determined by the judges.

After developing the scale as stated above, the respondents are asked during the administration of the scale to check the statements with which they agree. The median value of the statements that they check is worked out and this establishes their score or quantifies their opinion. It may be noted that in the actual instrument the statements are arranged in random order of scale value. If the values are valid and if the opinionnaire deals with only one attitude dimension, the typical respondent will choose one or several contiguous items (in terms of scale values) to reflect his views. However, at times divergence may occur when a statement appears to tap a different attitude dimension.

The Thurstone method has been widely used for developing differential scales which are utilised to measure attitudes towards varied issues like war, religion, etc. Such scales are considered most appropriate and reliable when used for measuring a single attitude. But an important deterrent to their use is the cost and effort required to develop them. Another weakness of such scales is that the values assigned to various statements by the judges may reflect their own attitudes. The method is not completely objective; it involves ultimately subjective decision process. Critics of this method also opine that some other scale designs give more information about the respondent's attitude in comparison to differential scales.

### *Summated Scales (or Likert-type Scales)*

Summated scales (or Likert-type scales) are developed by utilizing the item analysis approach wherein a particular item is evaluated on the basis of how well it discriminates between those persons whose total score is high and those whose score is low. Those items or statements that best meet this sort of discrimination test are included in the final instrument.

Thus, summated scales consist of a number of statements which express either a favourable or unfavourable attitude towards the given object to which the respondent is asked to react. The respondent indicates his agreement or disagreement with each statement in the instrument. Each response is given a numerical score, indicating its favourableness or unfavourableness, and the scores are totalled to measure the respondent's attitude. In other words, the overall score represents the respondent's position on the continuum of favourable-unfavourableness towards an issue.

Most frequently used summated scales in the study of social attitudes follow the pattern devised by Likert. For this reason they are often referred to as Likert-type scales. In a Likert scale, the respondent is asked to respond to each of the statements in terms of several degrees, usually five degrees (but at times 3 or 7 may also be used) of agreement or disagreement. For example, when asked to express opinion whether one considers his job quite pleasant, the respondent may respond in any one of the following ways: (i) strongly agree, (ii) agree, (iii) undecided, (iv) disagree, (v) strongly disagree.

We find that these five points constitute the scale. At one extreme of the scale there is strong agreement with the given statement and at the other, strong disagreement, and between them lie intermediate points. We may illustrate this as under:

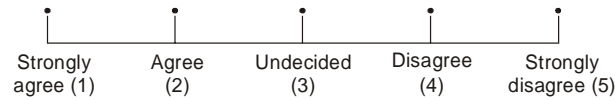


Fig. 5.3

Each point on the scale carries a score. Response indicating the least favourable degree of job satisfaction is given the least score (say 1) and the most favourable is given the highest score (say 5). These score—values are normally not printed on the instrument but are shown here just to indicate the scoring pattern. The Likert scaling technique, thus, assigns a scale value to each of the five responses. The same thing is done in respect of each and every statement in the instrument. This way the instrument yields a total score for each respondent, which would then measure the respondent's favourableness toward the given point of view. If the instrument consists of, say 30 statements, the following score values would be revealing.

$30 \times 5 = 150$  Most favourable response possible

$30 \times 3 = 90$  A neutral attitude

$30 \times 1 = 30$  Most unfavourable attitude.

The scores for any individual would fall between 30 and 150. If the score happens to be above 90, it shows favourable opinion to the given point of view, a score of below 90 would mean unfavourable opinion and a score of exactly 90 would be suggestive of a neutral attitude.

**Procedure:** The procedure for developing a Likert-type scale is as follows:

- (i) As a first step, the researcher collects a large number of statements which are relevant to the attitude being studied and each of the statements expresses definite favourableness or unfavourableness to a particular point of view or the attitude and that the number of favourable and unfavourable statements is approximately equal.
- (ii) After the statements have been gathered, a trial test should be administered to a number of subjects. In other words, a small group of people, from those who are going to be studied finally, are asked to indicate their response to each statement by checking one of the categories of agreement or disagreement using a five point scale as stated above.
- (iii) The response to various statements are scored in such a way that a response indicative of the most favourable attitude is given the highest score of 5 and that with the most unfavourable attitude is given the lowest score, say, of 1.
- (iv) Then the total score of each respondent is obtained by adding his scores that he received for separate statements.
- (v) The next step is to array these total scores and find out those statements which have a high discriminatory power. For this purpose, the researcher may select some part of the highest and the lowest total scores, say the top 25 per cent and the bottom 25 per cent. These two extreme groups are interpreted to represent the most favourable and the least favourable attitudes and are used as criterion groups by which to evaluate individual statements. This

way we determine which statements consistently correlate with low favourability and which with high favourability.

- (vi) Only those statements that correlate with the total test should be retained in the final instrument and all others must be discarded from it.

**Advantages:** The Likert-type scale has several advantages. Mention may be made of the important ones.

- (a) It is relatively easy to construct the Likert-type scale in comparison to Thurstone-type scale because Likert-type scale can be performed without a panel of judges.
- (b) Likert-type scale is considered more reliable because under it respondents answer each statement included in the instrument. As such it also provides more information and data than does the Thurstone-type scale.
- (c) Each statement, included in the Likert-type scale, is given an empirical test for discriminating ability and as such, unlike Thurstone-type scale, the Likert-type scale permits the use of statements that are not manifestly related (to have a direct relationship) to the attitude being studied.
- (d) Likert-type scale can easily be used in respondent-centred and stimulus-centred studies i.e., through it we can study how responses differ between people and how responses differ between stimuli.
- (e) Likert-type scale takes much less time to construct, it is frequently used by the students of opinion research. Moreover, it has been reported in various research studies\* that there is high degree of correlation between Likert-type scale and Thurstone-type scale.

**Limitations:** There are several limitations of the Likert-type scale as well. One important limitation is that, with this scale, we can simply examine whether respondents are more or less favourable to a topic, but we cannot tell how much more or less they are. There is no basis for belief that the five positions indicated on the scale are equally spaced. The interval between 'strongly agree' and 'agree', may not be equal to the interval between "agree" and "undecided". This means that Likert scale does not rise to a stature more than that of an ordinal scale, whereas the designers of Thurstone scale claim the Thurstone scale to be an interval scale. One further disadvantage is that often the total score of an individual respondent has little clear meaning since a given total score can be secured by a variety of answer patterns. It is unlikely that the respondent can validly react to a short statement on a printed form in the absence of real-life qualifying situations. Moreover, there "remains a possibility that people may answer according to what they think they should feel rather than how they do feel."<sup>4</sup> This particular weakness of the Likert-type scale is met by using a cumulative scale which we shall take up later in this chapter.

In spite of all the limitations, the Likert-type summated scales are regarded as the most useful in a situation wherein it is possible to compare the respondent's score with a distribution of scores from some well defined group. They are equally useful when we are concerned with a programme of

\*A.L. Edwards and K.C. Kenney, "A comparison of the Thurstone and Likert techniques of attitude scale construction", *Journal of Applied Psychology*, 30, 72–83, 1946.

<sup>4</sup>John W. Best and James V. Kahn, "Research in Education", 5 ed., Prentice-Hall of India Pvt. Ltd., New Delhi, 1986, p. 183.

change or improvement in which case we can use the scales to measure attitudes before and after the programme of change or improvement in order to assess whether our efforts have had the desired effects. We can as well correlate scores on the scale to other measures without any concern for the absolute value of what is favourable and what is unfavourable. All this accounts for the popularity of Likert-type scales in social studies relating to measuring of attitudes.

**Cumulative scales:** Cumulative scales or Louis Guttman's scalogram analysis, like other scales, consist of series of statements to which a respondent expresses his agreement or disagreement. The special feature of this type of scale is that statements in it form a cumulative series. This, in other words, means that the statements are related to one another in such a way that an individual, who replies favourably to say item No. 3, also replies favourably to items No. 2 and 1, and one who replies favourably to item No. 4 also replies favourably to items No. 3, 2 and 1, and so on. This being so an individual whose attitude is at a certain point in a cumulative scale will answer favourably all the items on one side of this point, and answer unfavourably all the items on the other side of this point. The individual's score is worked out by counting the number of points concerning the number of statements he answers favourably. If one knows this total score, one can estimate as to how a respondent has answered individual statements constituting cumulative scales. The major scale of this type of cumulative scales is the Guttman's scalogram. We attempt a brief description of the same below.

The technique developed by Louis Guttman is known as scalogram analysis, or at times simply 'scale analysis'. Scalogram analysis refers to the procedure for determining whether a set of items forms a unidimensional scale. A scale is said to be unidimensional if the responses fall into a pattern in which endorsement of the item reflecting the extreme position results also in endorsing all items which are less extreme. Under this technique, the respondents are asked to indicate in respect of each item whether they agree or disagree with it, and if these items form a unidimensional scale, the response pattern will be as under:

**Table 5.3:** Response Pattern in Scalogram Analysis

| <i>Item Number</i> |          |          |          | <i>Respondent Score</i> |
|--------------------|----------|----------|----------|-------------------------|
| <i>4</i>           | <i>3</i> | <i>2</i> | <i>1</i> |                         |
| X                  | X        | X        | X        | 4                       |
| –                  | X        | X        | X        | 3                       |
| –                  | –        | X        | X        | 2                       |
| –                  | –        | –        | X        | 1                       |
| –                  | –        | –        | –        | 0                       |

X = Agree

– = Disagree

A score of 4 means that the respondent is in agreement with all the statements which is indicative of the most favourable attitude. But a score of 3 would mean that the respondent is not agreeable to item 4, but he agrees with all others. In the same way one can interpret other values of the respondents' scores. This pattern reveals that the universe of content is scalable.

**Procedure:** The procedure for developing a scalogram can be outlined as under:

- The universe of content must be defined first of all. In other words, we must lay down in clear terms the issue we want to deal within our study.
- The next step is to develop a number of items relating the issue and to eliminate by inspection the items that are ambiguous, irrelevant or those that happen to be too extreme items.
- The third step consists in pre-testing the items to determine whether the issue at hand is scalable (The pretest, as suggested by Guttman, should include 12 or more items, while the final scale may have only 4 to 6 items. Similarly, the number of respondents in a pretest may be small, say 20 or 25 but final scale should involve relatively more respondents, say 100 or more).

In a pretest the respondents are asked to record their opinions on all selected items using a Likert-type 5-point scale, ranging from 'strongly agree' to 'strongly disagree'. The strongest favourable response is scored as 5, whereas the strongest unfavourable response as 1. The total score can thus range, if there are 15 items in all, from 75 for most favourable to 15 for the least favourable.

Respondent opinionnaires are then arrayed according to total score for analysis and evaluation. If the responses of an item form a cumulative scale, its response category scores should decrease in an orderly fashion as indicated in the above table. Failure to show the said decreasing pattern means that there is overlapping which shows that the item concerned is not a good cumulative scale item i.e., the item has more than one meaning. Sometimes the overlapping in category responses can be reduced by combining categories. After analysing the pretest results, a few items, say 5 items, may be chosen.

- The next step is again to total the scores for the various opinionnaires, and to rearrange them to reflect any shift in order, resulting from reducing the items, say, from 15 in pretest to, say, 5 for the final scale. The final pretest results may be tabulated in the form of a table given in Table 5.4.

**Table 5.4: The Final Pretest Results in a Scalogram Analysis\***

| Scale type  | 5            | 12 | Item<br>3 | 10 | 7 | Errors<br>per case | Number of<br>cases | Number of<br>errors |
|-------------|--------------|----|-----------|----|---|--------------------|--------------------|---------------------|
| 5 (perfect) | X            | X  | X         | X  | X | 0                  | 7                  | 0                   |
| 4 (perfect) | –            | X  | X         | X  | X | 0                  | 3                  | 0                   |
| (nonscale)  | –            | X  | –         | X  | X | 1                  | 1                  | 1                   |
| (nonscale)  | –            | X  | X         | –  | X | 1                  | 2                  | 2                   |
| 3 (perfect) | –            | –  | X         | X  | X | 0                  | 5                  | 0                   |
| 2 (perfect) | –            | –  | –         | X  | X | 0                  | 2                  | 0                   |
| 1 (perfect) | –            | –  | –         | –  | X | 0                  | 1                  | 0                   |
| (nonscale)  | –            | –  | X         | –  | – | 2                  | 1                  | 2                   |
| (nonscale)  | –            | –  | X         | –  | – | 2                  | 1                  | 2                   |
| 0 (perfect) | –            | –  | –         | –  | – | 0                  | 2                  | 0                   |
|             | <i>n</i> = 5 |    |           |    |   |                    | <i>N</i> = 25      | <i>e</i> = 7        |

\* (Figures in the table are arbitrary and have been used to explain the tabulation process only.)



The table shows that five items (numbering 5, 12, 3, 10 and 7) have been selected for the final scale. The number of respondents is 25 whose responses to various items have been tabulated along with the number of errors. Perfect scale types are those in which the respondent's answers fit the pattern that would be reproduced by using the person's total score as a guide. *Non-scale types* are those in which the category pattern differs from that expected from the respondent's total score i.e., non-scale cases have deviations from unidimensionality or errors. Whether the items (or series of statements) selected for final scale may be regarded a perfect cumulative (or a unidimensional scale), we have to examine on the basis of the coefficient of reproducibility. Guttman has set 0.9 as the level of minimum reproducibility in order to say that the scale meets the test of unidimensionality. He has given the following formula for measuring the level of reproducibility:

$$\text{Guttman's Coefficient of Reproducibility} = 1 - e/n(N)$$

where  $e$  = number of errors

$n$  = number of items

$N$  = number of cases

For the above table figures,

$$\text{Coefficient of Reproducibility} = 1 - 7/5(25) = .94$$

This shows that items number 5, 12, 3, 10 and 7 in this order constitute the cumulative or unidimensional scale, and with this we can reproduce the responses to each item, knowing only the total score of the respondent concerned.

Scalogram, analysis, like any other scaling technique, has several advantages as well as limitations. One advantage is that it assures that only a single dimension of attitude is being measured. Researcher's subjective judgement is not allowed to creep in the development of scale since the scale is determined by the replies of respondents. Then, we require only a small number of items that make such a scale easy to administer. Scalogram analysis can appropriately be used for personal, telephone or mail surveys. The main difficulty in using this scaling technique is that in practice perfect cumulative or unidimensional scales are very rarely found and we have only to use its approximation testing it through coefficient of reproducibility or examining it on the basis of some other criteria. This method is not a frequently used method for the simple reason that its development procedure is tedious and complex. Such scales hardly constitute a reliable basis for assessing attitudes of persons towards complex objects for predicting the behavioural responses of individuals towards such objects. Conceptually, this analysis is a bit more difficult in comparison to other scaling methods.

### Factor Scales\*

Factor scales are developed through factor analysis or on the basis of intercorrelations of items which indicate that a common factor accounts for the relationships between items. Factor scales are particularly "useful in uncovering latent attitude dimensions and approach scaling through the concept of multiple-dimension attribute space."<sup>5</sup> More specifically the two problems viz., how to deal

\* A detailed study of the factor scales and particularly the statistical procedures involved in developing factor scales is beyond the scope of this book. As such only an introductory idea of factor scales is presented here.

<sup>5</sup> C. William Emory, *Business Research Methods*, p. 264–65.

appropriately with the universe of content which is multi-dimensional and how to uncover underlying (latent) dimensions which have not been identified, are dealt with through factor scales. An important factor scale based on factor analysis is *Semantic Differential (S.D.)* and the other one is *Multidimensional Scaling*. We give below a brief account of these factor scales.

**Semantic differential scale:** Semantic differential scale or the S.D. scale developed by Charles E. Osgood, G.J. Suci and P.H. Tannenbaum (1957), is an attempt to measure the psychological meanings of an object to an individual. This scale is based on the presumption that an object can have different dimensions of connotative meanings which can be located in multidimensional property space, or what can be called the semantic space in the context of S.D. scale. This scaling consists of a set of bipolar rating scales, usually of 7 points, by which one or more respondents rate one or more concepts on each scale item. For instance, the S.D. scale items for analysing candidates for leadership position may be shown as under:

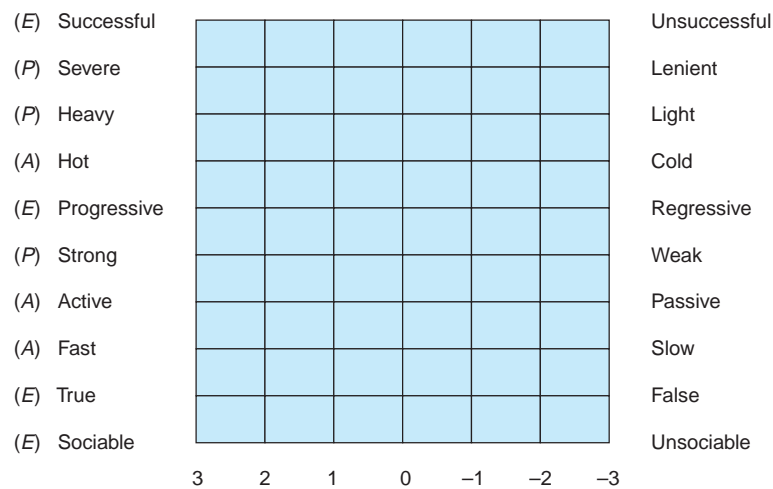


Fig. 5.4

Candidates for leadership position (along with the concept—the ‘ideal’ candidate) may be compared and we may score them from +3 to -3 on the basis of the above stated scales. (The letters, *E*, *P*, *A* showing the relevant factor viz., evaluation, potency and activity respectively, written along the left side are not written in actual scale. Similarly the numeric values shown are also not written in actual scale.)

Osgood and others did produce a list of some adjective pairs for attitude research purposes and concluded that semantic space is multidimensional rather than unidimensional. They made sincere efforts and ultimately found that three factors, viz., evaluation, potency and activity, contributed most to meaningful judgements by respondents. The evaluation dimension generally accounts for 1/2 and 3/4 of the extractable variance and the other two factors account for the balance.

**Procedure:** Various steps involved in developing S.D. scale are as follows:

- (a) First of all the concepts to be studied are selected. The concepts are usually chosen by personal judgement, keeping in view the nature of the problem.

- (b) The next step is to select the scales bearing in mind the criterion of factor composition and the criterion of scale's relevance to the concepts being judged (it is common practice to use at least three scales for each factor with the help of which an average factor score has to be worked out). One more criterion to be kept in view is that scales should be stable across subjects and concepts.
- (c) Then a panel of judges are used to rate the various stimuli (or objects) on the various selected scales and the responses of all judges would then be combined to determine the composite scaling.

To conclude, "the S.D. has a number of specific advantages. It is an efficient and easy way to secure attitudes from a large sample. These attitudes may be measured in both direction and intensity. The total set of responses provides a comprehensive picture of the meaning of an object, as well as a measure of the subject doing the rating. It is a standardised technique that is easily repeated, but escapes many of the problems of response distortion found with more direct methods."<sup>6</sup>

**Multidimensional scaling:** Multidimensional scaling (MDS) is relatively more complicated scaling device, but with this sort of scaling one can scale objects, individuals or both with a minimum of information. Multidimensional scaling (or MDS) can be characterized as a set of procedures for portraying perceptual or affective dimensions of substantive interest. It "provides useful methodology for portraying subjective judgements of diverse kinds."<sup>7</sup> MDS is used when all the variables (whether metric or non-metric) in a study are to be analyzed simultaneously and all such variables happen to be independent. The underlying assumption in MDS is that people (respondents) "perceive a set of objects as being more or less similar to one another on a number of dimensions (usually uncorrelated with one another) instead of only one."<sup>8</sup> Through MDS techniques one can represent geometrically the locations and interrelationships among a set of points. In fact, these techniques attempt to locate the points, given the information about a set of interpoint distances, in space of one or more dimensions such as to best summarise the information contained in the interpoint distances. The distances in the solution space then optimally reflect the distances contained in the input data. For instance, if objects, say  $X$  and  $Y$ , are thought of by the respondent as being most similar as compared to all other possible pairs of objects, MDS techniques will position objects  $X$  and  $Y$  in such a way that the distance between them in multidimensional space is shorter than that between any two other objects.

Two approaches, viz., the metric approach and the non-metric approach, are usually talked about in the context of MDS, while attempting to construct a space containing  $m$  points such that  $m(m-1)/2$  interpoint distances reflect the input data. The *metric approach to MDS* treats the input data as interval scale data and solves applying statistical methods for the additive constant\* which

<sup>6</sup> *Ibid.*, p. 260.

<sup>7</sup> Paul E. Green, "Analyzing Multivariate Data", p. 421.

<sup>8</sup> Jagdish N. Sheth, "The Multivariate Revolution in Marketing Research", quoted in "Marketing Research" by Danny N. Bellenger and Barnett A. Greenberg, p. 255.

\* Additive constant refers to that constant with which one can, either by subtracting or adding, convert interval scale to a ratio scale. For instance, suppose we know that distances, say  $a-b$ ,  $b-c$ ,  $c-d$  among stimuli on a ratio scale are 7, 6 and 3 respectively. If one were to subtract 3 from each of these distances, they would be 4, 3 and 0 respectively. The converted distances would be on an interval scale of measurement, but not on a ratio scale. Obviously, one can add 3 to all the converted distances and reach the ratio scale of distances. Thus 3 will be taken as the additive constant in this case. Well defined iterative approach is employed in practice for estimating appropriate additive constant.

minimises the dimensionality of the solution space. This approach utilises all the information in the data in obtaining a solution. The data (i.e., the metric similarities of the objects) are often obtained on a bipolar similarity scale on which pairs of objects are rated one at a time. If the data reflect exact distances between real objects in an  $r$ -dimensional space, their solution will reproduce the set of interpoint distances. But as the true and real data are rarely available, we require random and systematic procedures for obtaining a solution. Generally, the judged similarities among a set of objects are statistically transformed into distances by placing those objects in a multidimensional space of some dimensionality.

The *non-metric approach* first gathers the non-metric similarities by asking respondents to rank order all possible pairs that can be obtained from a set of objects. Such non-metric data is then transformed into some arbitrary metric space and then the solution is obtained by reducing the dimensionality. In other words, this non-metric approach seeks “a representation of points in a space of minimum dimensionality such that the rank order of the interpoint distances in the solution space maximally corresponds to that of the data. This is achieved by requiring only that the distances in the solution be monotone with the input data.”<sup>9</sup> The non-metric approach has come into prominence during the sixties with the coming into existence of high speed computers to generate metric solutions for ordinal input data.

The significance of MDS lies in the fact that it enables the researcher to study “the perceptual structure of a set of stimuli and the cognitive processes underlying the development of this structure. Psychologists, for example, employ multidimensional scaling techniques in an effort to scale psychophysical stimuli and to determine appropriate labels for the dimensions along which these stimuli vary.”<sup>10</sup> The MDS techniques, in fact, do away with the need in the data collection process to specify the attribute(s) along which the several brands, say of a particular product, may be compared as ultimately the MDS analysis itself reveals such attribute(s) that presumably underlie the expressed relative similarities among objects. Thus, MDS is an important tool in attitude measurement and the techniques falling under MDS promise “a great advance from a series of unidimensional measurements (e.g., a distribution of intensities of feeling towards single attribute such as colour, taste or a preference ranking with indeterminate intervals), to a perceptual mapping in multidimensional space of objects ... company images, advertisement brands, etc.”<sup>11</sup>

In spite of all the merits stated above, the MDS is not widely used because of the computation complications involved under it. Many of its methods are quite laborious in terms of both the collection of data and the subsequent analyses. However, some progress has been achieved (due to the pioneering efforts of Paul Green and his associates) during the last few years in the use of non-metric MDS in the context of market research problems. The techniques have been specifically applied in “finding out the perceptual dimensions, and the spacing of stimuli along these dimensions, that people, use in making judgements about the relative similarity of pairs of Stimuli.”<sup>12</sup> But, “in the long run, the worth of MDS will be determined by the extent to which it advances the behavioral sciences.”<sup>13</sup>

<sup>9</sup> Robert Ferber (ed.), *Handbook of Marketing Research*, p. 3–51.

<sup>10</sup> *Ibid.*, p. 3–52.

<sup>11</sup> G.B. Giles, *Marketing*, p. 43.

<sup>12</sup> Paul E. Green, *Analyzing Multivariate Data*, p. 421.

<sup>13</sup> Jum C. Nunnally, *Psychometric Theory*, p. 496.

### Questions

1. What is the meaning of measurement in research? What difference does it make whether we measure in terms of a nominal, ordinal, interval or ratio scale? Explain giving examples.
2. Are you in agreement with the following statements? If so, give reasons:
  - (1) Validity is more critical to measurement than reliability.
  - (2) Stability and equivalence aspects of reliability essentially mean the same thing.
  - (3) Content validity is the most difficult type of validity to determine.
  - (4) There is no difference between concept development and concept specification.
  - (5) Reliable measurement is necessarily a valid measurement.
3. Point out the possible sources of error in measurement. Describe the tests of sound measurement.
4. Are the following nominal, ordinal, interval or ratio data? Explain your answers.
  - (a) Temperatures measured on the Kelvin scale.
  - (b) Military ranks.
  - (c) Social security numbers.
  - (d) Number of passengers on buses from Delhi to Mumbai.
  - (e) Code numbers given to the religion of persons attempting suicide.
5. Discuss the relative merits and demerits of:
  - (a) Rating vs. Ranking scales.
  - (b) Summated vs. Cumulative scales.
  - (c) Scalogram analysis vs. Factor analysis.
6. The following table shows the results of a paired-comparison preference test of four cold drinks from a sample of 200 persons:

| <i>Name</i> | <i>Coca Cola</i> | <i>Limca</i> | <i>Goldspot</i> | <i>Thumps up</i> |
|-------------|------------------|--------------|-----------------|------------------|
| Coca Cola   | –                | 60*          | 105             | 45               |
| Limca       | 160              | –            | 150             | 70               |
| Goldspot    | 75               | 40           | –               | 65               |
| Thumps up   | 165              | 120          | 145             | –                |

\* To be read as 60 persons preferred Limca over Coca Cola.

- (a) How do these brands rank in overall preference in the given sample.
  - (b) Develop an interval scale for the four varieties of cold drinks.
7. (1) Narrate the procedure for developing a scalogram and illustrate the same by an example.  
 (2) Workout Guttman's coefficient of reproducibility from the following information:
 

Number of cases ( $N$ ) = 30  
 Number of items ( $n$ ) = 6  
 Number of errors ( $e$ ) = 10

Interpret the meaning of coefficient you work out in this example.
8. Write short notes on:
  - (a) Semantic differential scale;
  - (b) Scalogram analysis;

- (c) Likert-type scale;
  - (d) Arbitrary scales;
  - (e) Multidimensional scaling (MDS).
9. Describe the different methods of scale construction, pointing out the merits and demerits of each.
10. "Scaling describes the procedures by which numbers are assigned to various degrees of opinion, attitude and other concepts." Discuss. Also point out the bases for scale classification.