A REVIEW AND ANALYSIS
OF SCALING METHODS

Discussion of Scaling Methods:

Scaling Methods differ from the other methods discussed in Chapter 3. Scaling aims at the construction of a measuring device which will distribute individuals along a continuum running from one extremity of response through neutral to the opposite extremity of response. This continuum represents the entire range from the positive aspects to the negative aspects of an attitude variable. On an attitude scale this range is divided into measurable units to which numerical values are assigned, to indicate a person's position on a given issue.

The typical attitude scale consists of a series of short and carefully formulated statements or propositions dealing with several selected aspects of issues, institutions or groups. The items are so selected as to cover the entire range from the most favourable (positive) to the least favourable (negative) responses. The individual is required to react verbally with expressions of approval (agreement) or disapproval (disagreement) to each of the given set of items. Attitudes on a given issue are then inferred by summation of reactions to the set of items. Individual's position on that issue is, thus determined along a quantitative scale continuum. Scaling is
considered to be the process of determining benchmarks on the yardstick. Measurement consists of applying this yardstick to evaluate the objects (Angoff 1960).

Types of Scales: Scales differ markedly in type and method of construction, though their objective is same. Four basic types of scales, upon which the various scaling methods are based, have been distinguished by Stevens (1961).

1. Nominal Scales: This is the simplest form of scale, consisting of two or more named categories in which objects or individuals or responses are classified. These various categories are mutually exclusive. Groups of objects either clearly belong or do not belong to the same class. The numbers or name assigned to the objects in a category serves as a label. Any other set of labels may be as adequate. For example, classification of individuals according to nationalities constitutes a nominal scale. Such a scale is used in exploratory phases.

2. Ordinal Scales: When objects can be arranged in a rank order, they are said to form an ordinal scale. It defines the relative position of objects or individuals with respect to a trait with no implications as to the distance between positions.
3. Interval Scales: When the intervals between objects on a scale can be measured then an interval scale exists. It indicates the exact distance between the two stimuli.

4. Ratio Scales: It is an interval scale with the addition that the origin is an absolute zero, not an arbitrary zero.

In addition to the four types of scales mentioned above (Coombs 1953), describes two more scales. He adds a "Partially ordered scale" between the Stevens' (1951) hierarchy of nominal and ordinal scales. According to Coombs, the objects may be simply classified, partially ordered or completely ordered. A partially ordered scale would result if the objects could be ordered with respect to some attributes but could not be ordered with respect to some other attributes. Sometimes the objects in one class are more than just different from the objects of another class. They are neither totally different nor totally the same. Whenever an attribute is made-up of two or more attributes which do not combine additively a partially ordered scale results.

Another type of scale discussed by Coombs is the "Ordered Metric Scale", which lies between the ordinal and interval scales of Steven's hierarchy. In this type of
scale the objects are ordered and also the intervals between the objects are ordered. It indicates the relation as well as the differences or distances between the objects. The interval scale indicates the exact amount of the magnitude of differences between the objects, whereas the ordered metric scale distinguishes the differences in terms of "greater" or "less". Thus in Coombs' hierarchy there are six types of scales: (i) nominal scale - indicates no relation between the stimuli, (ii) partially ordered scale - indicates partial relationship, not absolute, (iii) ordinal scale - indicates absolute relationship, for example "A is greater than B", "B is greater than C" etc., (iv) ordered metric scale - indicates beside the absolute relationship between the stimuli a partial or crude relationship between the differences among the stimuli, for example "the distance between A and B is 'more' or 'less' than the distance between B and C", (v) interval scale - indicates an exact relationship between the differences, for example, "A is twice or thrice as great than B as B is to C" (vi) ratio scale - indicates in addition the absolute beginning.

**Attitude Scaling Methods**: In social sciences, there are nominal, ordinal and even interval scales. But to achieve the precision of a ratio scale, the establishment of absolute zero, is difficult to obtain. Some procedures have been devised to enable the social scientists, psychometricians and psychophysicists to
determine the scale values of a series of objects, events, etc., with respect to some attribute. The various scaling methods and procedures have been classified by different criteria. Torgerson (1958) and Gullickson and Kessick (1960) discuss the various ways in which these methods have been classified.

We are here concerned only with those scaling methods which are used in attitude measurement.

The early efforts of attitude measurement were limited to putting before the respondents a battery of questions, selected on an a priori basis. Numerical values were assigned arbitrarily. The procedures of standardisation were not known. It is well known that first major step in the movement of attitude measurement was taken by Thurstone in 1928 when he first adapted the method of psychophysical scaling originated by Fechner to the scaling of judgment of favourableness - unfavourableness toward various objects. Ever since, there has been an increasing realisation of the need of standardised attitude scales. Due emphasis has been put on developing methods to scale the attitude items on a continuum.

Two general types of approaches that have been widely used in attitude scale construction have been distinguished by Edwards (1957). One of these consists of asking a
group of judges to evaluate the given items. This is exemplified by the Thurstone (Thurstone and Chave 1929) method of equal-appearing-intervals. The second method of developing attitude scales is based upon the subjects' responses to attitude statements, for example, the Likert (1932) method of summed ratings.

Attitude scales can be variously classified depending upon the basis selected for classification as illustrated below:

Remmers (1954) distinguishes six types of scales:

1. Apriori scales, which is the crudest method of measuring attitudes. In these, Remmers includes case method, ballot counting, Bogardus (1925 a, b) method etc. 2. Psychophysical scales, as for example, Thurstone's method. 3. Sigma scales, as for example, Likert's (1932) method. 4. Master scales, like Remmers and Silance (1934) master scale. 5. Behaviour scales, as Rosander's behaviour scale and 6. Guttman's unidimensional scale.

Doby et al. (1954) mention three types of scales:

2. Judgement Scales: (a) arbitrary scales; it contains nothing more than the arbitrary summation of responses to a series of questions or items. It does not tell about the underlying structure of items; (b) Consensus scales, by the technique items are rated by a group of judges.

2. Complex Operational Scales: (a) cumulative scales, provide a test of unidimensionality; (b) Item-analysis techniques, items are selected depending upon their
relationship to the total score of the scale or upon
their discriminatory power to differentiate between
the high and low groups.

3. Factor Scales: Intercorrelations between the items
are worked and factors are extracted.

Young (1956) classifies the scales in the following
manner: 1. arbitrary scales, 2. scales in which the
item scale values are determined by a panel of judges,
3. scales based on item analysis and, 4. scales
constructed by scale-analysis techniques.

Krach et al. (1962) distinguish five principle methods
of attitude scale construction: 1. method of equal-
appearing intervals, 2. method of summated ratings,
3. the social distance scale, 4. cumulative scale and,
5. scale discrimination technique. Klineberg (1954)
differentiates the following types of scales:
1. a priori scale, the scoring is arbitrary, depending
upon the judgment of the investigator, for example,
Bogards (1925 a, b) social distance scale; 2.
Psychophysical scales, Likert and 4. Guttman type scales.

The following scaling methods have been discussed by
Gréen (1954). 1. the Judgment methods, in which the
scales are developed from judges' ratings of the items,
for example, method of equal-appearing-intervals,
method of paired comparisons and method of successive intervals, 2. the response methods in which scales are developed from the data given by the respondents, for example, the method of summed ratings, scalogram analysis, the unfolding technique and the latent structure analysis, 3. rating scales.


Besides the writers mentioned above, Nunally (1955), Ferguson (1955), Albig (1956), Britt (1956), Sells (1960), Angoff (1960), McNewar (1960), Guilford (1959), Dunlap and Kroll (1939) are, among the many others who classify the various scales and scaling methods in their own ways. The classifications have been briefly noted above.

The methods of scaling are discussed below:

1. Method of paired comparisons: The stimuli is presented to the subjects in pairs and he has to judge which of the two is preferable. This method was originated by Fechner and first applied to attitude measurement by Thurstone (1927). The stimuli thus presented to the subject consists of n (n-1)/2 pairs. For example, a list of 5 nationalities may be presented to the subject in all possible pairs, i.e., by pairing each nationality
with every other nationality. There will be $5(5-1) = 10$
pairs of nationalities and the subject may be asked to
underline the nationality he prefers. Instead of
nationalities, pairs of attitude statements may be
presented to the subjects. Analysis of the data
and observation of scale values is then done by
adopting certain standard procedures of calculation
(see Edwards 1957).

The greatest disadvantage of this method is that it can
be used only when the stimuli to be scaled is not too
large. For instance, if the number of items is 30,
the number of comparisons would be $30(30-1)/2 = 435$,
which may become unmanageable. Also, the procedure is
too cumbersome.

2. Method of equal-appearing-intervals: One of the
earliest and still most widely used method of constructing
attitude scales was developed by Thurstone and his colleagues
in 1929 (Thurstone and Chave 1929). This method is often
referred as the "Thurstone technique". The main feature
of this method is the use of judges to determine the
points on the attitude continuum. The procedure used
in the method of equal-appearing-intervals is as follows:

First of all, the area of inquiry, for example, attitude
towards Church, war, religion, etc, is carefully
defined. The next step consists of collecting a
large number of statements expressive of attitudes
towards the specific issue concerned. These statements
should cover all the possible gradations and shades of opinion on the issue under study. Items are collected from various sources, like books, periodicals, newspapers, other literature, lectures, discussions and previous scales available on the same issue. These are then edited and carefully scrutinised so as to eliminate those found to be double-barrelled, ambiguous or not strictly pertinent to the attitude variable.

The edited statements are then mimeographed or printed on separate cards, one item on each card. This collection of statements is given to a large group of judges. Each judge is required to sort the items into 11 piles, representing various degrees of favourable and unfavourable attitudes. These piles are supposed to be equally spaced and are serially arranged and marked from A to K. The two extreme piles, A and K, are described as representing the most favourable and most unfavourable attitudes respectively. The middle-most pile F, is described to represent the neutral attitude. The rest of the points on the continuum, the positions intermediate between these points are not descriptively labeled. They are simply mentioned as expressing various degrees of favourable - unfavourable attitudes. The judges
are precisely instructed to sort the given items in the given piles on the basis of degrees of favourable and unfavourable attitudes expressed by the statements.

Distributions are then tabulated for each statement showing the frequency with which it was placed in each of the 11 piles. Two statistics are required for each of these distributions: the median and the interquartile range of $Q$. Medians and $Q$ values can be obtained by both mathematical or graphical methods. Thurstone and Chave (1929) however, recommend the graphical method. These two statistics obtained serve as the criteria for selecting statements to build up the final scale. The median is taken as the objective scale value of the item, and $Q$ serves as the measure of the degree or extent of ambiguity. Small $Q$ values indicate closer agreement among the judges regarding the placement of the item on the continuum, therefore, less ambiguity. A large $Q$ value is indicative of wider dispersion or less agreement and, therefore, greater ambiguity.

The attitude scale is built up by choosing a number of items which have low $Q$ values and whose scale values cover
the entire continuum as evenly as possible. In other words, the items to comprise the final scale are so selected as to show a uniform spread along the entire continuum. The final scale is, thus, made of more or less equi-distant-items. The assumption of equal-appearing-intervals among the items of the scale is basic in this method. The attitude scale is now ready for use. In practice, two parallel forms are prepared, each having 20 to 22 items, for purposes of reliability determination.

While measuring the attitudes, the subjects are instructed to endorse the statements which express their attitudes. The mean or median value of the endorsed statements is taken to be his attitude score.

Thurstone and Chave (1929) mention another criterion in addition to Q values for rejecting the irrelevant items. This is known as the criterion of irrelevance. But this is not much used in practice.

Various modifications have been introduced in the original Thurstone method for getting item ratings and deriving scale and Q values.
Ballin and Fasnessworth (1941) recommend a graphic continuum of rating by providing an 11 inch line to replace the sorting procedure. Along with the numbered list of statements, a sheet bearing 11 inch line is presented to the subjects. Farnsworth (1945) suggests yet another improvement by providing separate graphic lines for each item, instead of one line for all the statements.

Seashore and Havner (1933) introduced yet another variation. They provide the statements printed in booklets rather than separate slips and numbers from 1 to 11 are printed on the left of each statement. Subjects are to indicate their ratings by encircling the left hand numbers. This does away with the method of separate slips and piles.

Similarly, for computing medians and Q values Edwards and Kilpatrick (1948) recommend a large size master graph taped on to a ground glass plate illuminated from below. With the help of tracing papers ogive curves can be quickly drawn. Jurgenson (1943) has also prepared a nomograph for rapid determination of medians and Q values.
Thurstone method has been criticised for "end effect", influences of judges' own attitudes on ratings, etc. The assumption of equal-appearing-intervals has also been challenged.

One of the major criticisms against the method of equal-appearing-intervals is that the attitudes of Judges may influence their judgements. Studies have been conducted to investigate this assumption. Thurstone and his colleagues hold that the scale values of the statements are independent of the attitudes of judges. Hinkley (1932), Ferguson (1935), Pintner and Forlano (1937) and Prothro (1955) used different groups of judges with differing attitudes to sort the statements. The above writers found no influence of judges' personal attitudes on the scale values. Novland and Sheriff (1952) criticise these studies as not having included Judges with extreme attitudes. They also point out certain other inadequacies of Hinkley (1932) study. Using the same 114 statements as those used by Hinkley, Novland and Sheriff (1952) found results contrary to Hinkley study. Therefore, the latter writers concluded that attitudes of judges do affect the scale values of items. Novland and Sheriff cite various other studies
also which are in conflict with the Hinkley results.

Kelley et. al. (1955) after reviewing Hinkley (1932) and Hovland and Sherif (1952) studies conducted their own study by using the method of paired comparisons and concluded that the displacement effect of scale values due to attitudes of judges can be eliminated if the method of paired comparisons rather than of equal-appearing-intervals is used.

The question of the independance of scale values and attitudes of judges has, however, not yet been resolved one way or the other.

Another criticism against Thurstone method is that the judges may not regard the scale steps as equal-distant, which is the basis of this method. Farnsworth (1945) on an investigation found that many judges do not think in terms of equal units. But a part of this difficulty might be eliminated by giving clear instructions to judges.

Farnsworth (1943) points out that the scale values obtained by Thurstone method may change after a lapse of time. He compared the scale values of Thurstone Peterson scale toward war obtained in 1930-31 with...
those obtained in 1940-41 and found a significant change in item weights.

Fehrer (1982) argues that the judgment of an item in a scale depends upon the other items constituting the series. For example, the judgment of heaviness of a weight will depend not only on the absolute magnitude of the weight itself but also on the other weights in the series. If the weight is moved to a different series, the judgment of its heaviness will change. In an experiment, Fehrer (1952) found that the judgments of social-attitude scale statements are also likewise influenced by the particular series of items composing the scale. Moving the items into another series of items changes its scale value. However, extreme items do not show as much change as moderate items.

Many writers criticize Thurstone scaling method for the "end effect". Judges tend to chose more often the extreme categories. Some authors, however, maintain that the end effect which affects the distribution of the stimuli near the extremes can be
eliminated by providing extra categories.

Thus, the Thrustone technique has been variously criti­
icised but, modifications and methods have also been
suggested at the same time, to overcome these
weaknesses.

2. The method of successive intervals: This method
is used when there are too many items to be scaled,
not feasible by the method of paired comparisons.
The method of successive intervals was designed by
Thurstone and first reported by Saffir (1937). It
was designed to overcome some of the difficulties inherent
in the method of equal-appearing-intervals. Like the
method of equal-appearing-intervals, this method also
required a single judgment from each subject for each
statement to be scaled. The instructions to the
judges are similar to those given in the method of
equal-appearing-intervals. Same number of categories,
7, 9, or 11 are used for sorting the items. For
each statement a frequency distribution is obtained
showing the number of times the statement has been
placed in the given categories of favourable-unfavourable;
Cumulative frequencies are then obtained which are
expressed as cumulative p-roporions. The essential
difference of this method from the method of equal-appearing-intervals lies in the method of treating the data. The scaling problem in the method of successive intervals is to determine the estimates of the widths or the intervals making up the psychological continuum.

It has been found that the relationship between the two sets of scale values, obtained by the method of equal-appearing-intervals and paired comparisons, is approximately linear except at the two extremes. A statement having extreme scale value by the method of paired comparisons is shown to have less extreme scale value when scaled by the method of equal-appearing-intervals, that is, extreme statements shift toward the centre of equal-appearing-interval continuum. The method of successive intervals is designed to eliminate this weakness. It retains the simplicity of the method of equal-appearing-intervals and, at the same time, yields scale values that are linearly related to those obtained by the method of paired comparisons, over the entire range. The procedures of obtaining scale values are described by Edwards (1957), Green (1954) and Saffir (1937).
4. The method of summated ratings: This method was designed by R. Likert in 1932 and so has also come to be known as the "Likert technique". The method of summated ratings also starts with the collection of a large number of positive and negative statements about the attitude object. Approximately half the statements express favourable and half unfavourable attitudes. Unlike Thurstone approach, judges are not employed in this method. Instead, the scale is derived by item-analysis techniques. The items are presented to the subjects in the form of a questionnaire or attitude test. Each item is followed by multiple response categories. Likert (1932) used five response categories: strongly agree, agree, undecided, disagree and strongly disagree. The subjects indicate their attitude by checking the response which most nearly expresses their feelings on that item. The scoring can be done by two methods - "sigma-deviate scoring" and "arbitrary scoring". The former method involves scoring on the basis of normal curve transformations of the proportion of respondents choosing each category. Likert soon found that the
simple "arbitrary scoring" methods correlate .99 with the "sigma" method. Therefore, the latter method was abandoned in favour of 'arbitrary scoring' method. Arbitrary scoring method which is most commonly and widely used, consists of assigning arbitrary weights of 5, 4, 3, 2 and 1 to the five response categories in such a way that the highest weight is always assigned to the response that tends toward the favourable end, while the lowest weight is assigned to the response that tends towards the unfavourable end. A total score for each individual is obtained by summing his scores for the individual items.

Subsequently, item-analysis is carried out to select the most discriminating or valid items. Item-analysis data is secured in the following manner: 1. correlation of each item with the total score is computed. An item to be valid should have a high positive correlation with the total score. This reveals the internal consistency of the test. 2. The second method is to select the most discriminating items, that is, those items which discriminate between the
high and low scores. For comparing high and low groups on the given items, some people use 't' test, some use phi-coefficients, while some recommend the simple ordering of the magnitude of differences between the mean scores of high and low groups. The purpose is to select the most discriminating items by whatever method. Criterion groups consisting of 10, 27 or some other percent of the subjects in terms of their total scores are selected and compared to find out whether the individual items differentiate between the two groups. Thus, the criterion of internal consistency or discriminatory power helps the selection of the valid items for the final scale, which usually consists of 20 to 28 items.

Comparison of Thurstone and Likert methods: Questions have been raised as to which of the two methods is superior, Thurstone or Likert.

Ferguson (1941) reports that Likert obtained higher reliability coefficient on his method of scale construction than the one obtained by Thurstone and, therefore, concluded that his method is superior as it does away with the need of a judging group. But Ferguson (1941) criticises Likert as having used a scale
already constructed by Thurston technique. This may account for the higher reliability he obtained, because as Ferguson (1941) says "increasing the number of steps in psychological scaling increases reliability." Ferguson (1941) therefore, used four scales constructed by Likert technique and got them scaled by Thurstone method. But this design again does not give a fair test of the two methods. It can only find out how the Likert type selected items fall on Thurstone continuum. He found that items selected by Likert method, when scaled by Thurstone technique, fail to spread evenly over the scale continuum of Thurstone; they largely fall at the two ends. Ferguson (1941) therefore, concluded that Likert technique does not obviate the need of a judging group. But, here again, Ferguson committed a similar mistake for which he criticised Likert. He did not give Thurstone method a fair trial. Whereas, Likert in his study used the statements already scaled by Thurstone method, Ferguson limited the Thurstone scale to the statements already selected by Likert technique.

The valid comparison of the two methods can only be made
when the two scales are independently constructed.

One must have one original set of items, not items already sifted by Thurston or Likert procedures.

Edwards and Kenway (1946) conducted another investigation using an independent set of items and getting these items scaled independently by Likert and Thurston procedures. The reliability coefficients for the Likert type scale of 25 items was .94 and for the two parallel forms of Thurstone scales of 20 items each was .88. The correlation coefficient between Likert scale and Form A of Thurstone scale was .79 and Likert scale and Form B of Thurstone scale was .92. The writers concluded that the coefficients are sufficiently high to establish the fact that both methods will yield comparable results. Edwards and Kenney (1946) further say that scales can be constructed by the method of summated ratings more quickly and with less labour than by the method of equal-appearing-intervals. It has also been indicated by some writers (Ferguson (1941) quotes Hall) that Likert type scales give higher reliabilities with even fewer items.

Nunnally (1959) points out that Likert approach is more empirical as it deals directly with the respondents.
Moreover, the use of five point scale provides more information than a simple agree-disagree response. It better represents the finer shadings of attitudes and shows the extent of agreement and disagreement with each specific opinion. It offers insight into the intensity of belief or disbelief.

Thurstone method is superior in that it gives direct meaning to the attitude scores. Likert type scales are relative to the group. It can be interpreted adequately if norms are available, but large scale norms are seldom obtained. Remmers (1954) also points out that the Thurstone scale procedure gives absolute meaning to a scale score, whereas Likert type scale scores can only be interpreted in terms of norms.

5. Scalogram Analysis: A new departure in attitude measurement was provided by Guttman. This method is also known as "Guttman method". Guttman expressed the desirability of having a unidimensional scale. The main assumption in this method is that if we know the total score of an individual, we should be able to predict his response on the individual item. In a perfect Guttman scale, the total score of an
individual will have one-to-one relationship with the pattern of his responses to the items making up the scale. Knowing the total score of an individual, it would be possible to reproduce perfectly his responses to each of the items. If, for example, an individual endorses a very extreme item, he should endorse all the less extreme items in the scale.

Guttman strongly emphasizes that a scale should be strictly unidimensional. The test score should be a measure of one factor only. If a single quantitative score is to represent without ambiguity, the behaviour of an individual on a group of items, then knowing the individual's score, we must be able to know his behaviour on each and every item of the group, because each and every item is essentially designed to measure the same thing. Guttman calls this the principle of "reproducibility", that is, by knowing his total score, we should be able to reproduce his behaviour on each item. For example, let us suppose there are five sticks arranged in order of length, 'A' being the longest stick and 'E' being the shortest one. Now, we use these sticks to measure the heights of persons, we may stand each stick alongside each individual and record whether he is taller or shorter than the stick. If he is taller,
we record a plus sign and if shorter we record a minus sign. A person getting plus mark on stick 'A' will get plus mark on rest of the sticks. Another person receiving plus on stick 'C' and minus on stick 'B' should receive minus also on stick 'A' and plus in sticks 'D' and 'E'. Thus, knowing the total score we should be able to predict his markings on all the sticks. For example, a person getting 4 plus marks must have got minus on stick 'A' and plus on rest of the sticks and a person getting 3 plus signs must have got minus on 'A' and 'B' and plus on rest. In a similar manner, according to Guttman, responses should be predictable on attitude test item.

The attitude test data is treated by way of a scalogram. The items are ranked in order of ascending and descending positions and the individuals are also ranked in order of descending positions. The resulting pattern of responses forms a scalogram. A person endorsing the most extreme positions should be characterised by having endorsed the less extreme positions. This scale is also called "cumulative scale". The cumulative nature of scale is indicated by the fact that people in higher positions have
all the characteristics of people in lower scale positions, plus at least one additional positive response. The items have a cumulative property.

In a perfect Guttman scale an individual score bears one-to-one relationship with his response pattern, that is, his item responses are specified by his score. This method is an attempt to evaluate a set of statements to determine whether they meet the requirements of unidimensionality. In Guttman's words "we shall call a set of items of common content a scale if a person with higher rank than another person is just as high or higher on every item than the other persons". (Krech et al. 1962 - P.154).

The Guttman scaling procedure consists of defining an area of content, or the universe of attributes and then selecting a sample of statements to represent adequately the various aspects of this universe of content. Then, we must define a population of individuals whose behaviour with respect to the universe of content, we are interested in observing and so we obtain a sample of individuals. The behaviour of our sample of individuals now affords the means of testing the hypothesis of unidimensionality.
or reproducibility forwarded by Guttman.

There are several techniques for performing scale-analysis. Reamers (1954) mentions four such techniques.
1. least-squares principle of minimising errors of prediction; 2. the Scalogram board procedure,
3. the frequency tabulation technique and, 4. the Cornell technique. Edwards (1957) and Green (1954) also describe methods of conducting scale analysis.

Several points have been raised regarding the validity and applicability of this method. It is considered that the items obtained by this technique will be highly homogenous in content. That is why Ferguson (1952) argues "why is more than one item needed in a scale (p. 137)". In Guttman technique we only succeed in asking the same question in two different manners. Secondly, Ferguson argues that Guttman technique aims to demonstrate that there is a perfect correlation between each item and total score. Therefore, Guttman has done a little more than to set up an elaborate system to determine the correlation between an item and total score. This result can be obtained just as well by computation of a biserial,
The third argument of Ferguson (1952) is as to how to select the initial set of statements which will meet the Guttman criterion of scalability. Guttman asserts that the selection of a sample of statements is a matter of intuition and experience. Thurstone and Likert employ rigorous procedures to select the scalable items from an originally large number of items. In Guttman technique, Ferguson argues, one investigator may just happen to select the items which will meet Guttman requirements but another investigator may be less fortunate in his selection.

How the statements are selected remains a mystery (Edwards 1957). Edwards terms Guttman technique as a procedure for evaluating a set of statements, as to whether they meet the particular requirements. It is not a method for constructing attitude scales.

REMERS (1954) also argues that the entirely subjective nature of the selection of the original sample of items from the "content universe", for testing the hypothesis of scalability, makes it a focus of criticism.
Some practical drawbacks of the Guttman method have been pointed out as follows (Nunnally 1967). Almost no collection of statements will meet this criterion in practical work. Individual items are notoriously heavy with measurement error. The items we obtain as scalable may be so closely related in content as to constitute a near rewording of the same statements. They may have a very narrow range of content. Moreover, when the Guttman criterion is not met perfectly, the whole set of items has to be abandoned. There is little that can be done statistically to determine the underlying dimensions which prevent the scaling of items on a unidimensional basis.

Guttman technique has also been further criticised (Sherif and Sherif 1966). The analysis proceeds on an empirical and trial and error basis. Moreover, this type of scale becomes highly stereotyped. In practice no perfect scales are found.

Unidimensional scales are no doubt desirable but,
practically difficult to achieve (Krech and Crutchfield) (1948). No single item can be expected to reflect only one single attitude. For example, the judgment of approval or disapproval of an item, "the Government should provide free medical service for all people" may be determined, simultaneously, by attitudes about socialism, bureaucracy or humanitarianism. This item might find a place in a scale devoted to the measurement of any of these attitudes." There is nothing intrinsically improper in having an attitude test cover several different aspects of the attitude or belief, providing these aspects are parts of a constellation making up the belief or attitude" (Krech and Crutchfield 1948, P. 231). No one would object to a measurement of physical height of people that depends upon the number of inches from head to toe, yet, this total height is made up of heterogenous parts, the length of the leg, the neck, etc. What is to be avoided in a scale is the inclusion of irrelevant items.

"Reproducibility" is not an essential aspect of
measurement, is also emphasized by Doby et al. (1954). To quote the above example again the height of a person is made up of the lengths of the head, the neck, the trunk and the legs. The sum of all these parts would give the total height. But the knowledge of the total height does not imply a knowledge of the lengths of all these separate parts.

Any scale which is unidimensional will be composed of homogenous items, (Conrad 1946). Some heterogeneity among items is, however, tolerable. Because of the complex origin and complicated nature of many opinions and attitudes, strictly unidimensional scales in these fields are not possible to construct, except possible for the issues which are simple and have a narrow range. Unidimensionality is a desirable goal but if it is achieved by restricting the construction of scales to simple and narrow issues, than the ideal is achieved at the cost of social usefulness.

To predict average academic achievement, there is no harm in combining the high school grades, intelligence scores, etc., of a student, because
though the two refer to different aspects, they serve the same purpose of predicting the academic achievement. Similarly, a salesman's financial gain from one type of goods is entirely independent from his gain from another type of goods, but the two types of gains can be combined, to find out as to how much money he makes as a salesman. In a similar manner, responses to different items should be combined, if they have the common characteristic of pertinence to the attitude under consideration.

We are often confronted with a multidimensional phanomena and, there will be no harm in combining the allied aspects to form one phase. It must be carefully borne in mind that perfect unidimensionality and reproducibility in the field of attitudes can not be realised.

6. Scale-discrimination technique: The most widely and popularly discussed methods in the construction of attitude scales are the three: the "Thurstone technique", the "Likert technique", and the "Guttman technique", which have already been discussed. Each has its own advantages and disadvantages.
Edwards and Kilpatrick (1948) attempted to synthesize these three methods in one, in order to eliminate the weaknesses of each and retain only the strong points. This method also increases the steps in psychologically scaling, which according to Ferguson (1941) increases the reliability of the scale. The scale-discrimination technique is of comparatively recent origin. It was devised by Edwards and Kilpatrick in 1948.

Edwards (1957) points out that the difficulty encountered in Guttman method is the selection of the initial set of statements. Guttman leaves it to intuition and experience. Item-analysis and the method of equal-appearing-intervals have something to contribute at this stage. These methods may help in the selection of the set of initial statements.

The chief unsolved problem in the Thurstone method is the selection of the most discriminating items from the same scale interval. Edwards (1957) notes that items with approximately the same scale values may vary markedly in their discriminating power. He reports items within the same scale-interval,
having phi-values ranging from .24 to .78.

In Thurstone technique, there is no procedure for selecting the items with comparable \( q \) values, falling within the same scale interval but, which may differ markedly in their power to differentiate between the high and low criterion groups. Item-analysis or Likert technique will be of help at this stage.

The second difficulty in Thurstone technique is that the frequency or probability of endorsement of an item is related to the scale value of the items. Assuming a normal distribution of attitudes and permitting an agree and disagree response, we would expect the model frequencies of the items scaled at the extremes to be high. As we move in toward the centre of the continuum, we would expect the frequencies to be more evenly distributed between the two categories of response. This, however, would give a rigorous test of scalability. Because the inclusion of items with a .5 and .5 division of response along with the marginals will keep the
the coefficient of reproducibility from being spuriously high. But the difficulty here is that if we have a very homogenous group, say a group strongly opposed to capital punishment, then we would expect a majority of subjects to agree with the statement against capital punishment and some subjects to disagree with the statement favouring capital punishment. Thus, the modal frequencies for all the items will be very high and the coefficient of reproducibility may not be much larger than the minimum established. (the minimum coefficient of reproducibility is the average of modal frequencies for all the items).

The third difficulty is that the neutral items in Thurstone scale have poorer reproducibility. The neutral items are less discriminating, there being more or less equal probability of endorsement by individuals from opposite ends of continuum. Because of the probable overlap in the responses to these items of those with high and low rank-order scores, responses to these statements will not be reproducible from the rank-order scores. These items
These items do not show clear cutting points and hence are expected to contribute greatly to error and the coefficient of reproducibility will be decreased.

It has also been proved (Edwards and Kilpatric 1948) that the reproducibility of an item is related to its discriminatory power. More discriminating an item is higher the reproducibility. Likert technique helps us, at this point, to select the most discriminating items from the given set.

From these findings they reasoned out a combination of these scaling methods. The Thurstone technique will help select the initial set of statements. Likert technique will help retain the most discriminating ones and such a combination would enable one to select a relatively small set of statements, such as will have a good chance of meeting the Guttman requirements.

The initial step in scale-discrimination technique is similar to that followed by Thurstone or Likert. A large number of statements relating to the attitude object are collected and edited in accordance with
certain criteria. This collection of items is then subjected to Thurstone scaling procedure.

The assumption of equal-intervals is not vital to scale-discrimination technique. The medians (scale values) and Q values are obtained for each statement. The median of the Q values of all items is then calculated. Items falling above the median Q value are rejected as the ambiguous ones. This amounts to eliminating approximately 50 percent of statements.

The remaining 50 percent of items with the least degree of ambiguity are then drawn up in a Likert type scale. Each statement is followed by multiple choice response categories. Edwards and Kilpatrick (1940) used 6 categories, strongly agree, agree, mildly agree, mildly disagree, disagree, and strongly disagree. A fewer categories could also be used. The questionnaire is given to a new group of 200 to 300 subjects who are asked to respond to the items in terms of the response categories. The responses are scored by the method of summated ratings.
A total score for each individual is obtained. The discriminating power of each item can be calculated by any of several procedures. Edwards and Kilpatrick (1948) used Phi-coefficient because of its simplicity. It involves less computation and labour than most other procedures. The writers selected the top and bottom 27 percent, in terms of the total scores, as the "high" and "low" groups. A distribution is plotted for each statement, showing the frequency for each response category, separately for the two groups. Response categories are then dichotomised. The rule followed in combining the response categories was to tabulate the response categories for each item for the two groups and then to draw a line between the categories in such a manner so as to minimise the total number of subjects in the low group above the line and in the high group below the line. Details regarding this are given in Edwards (1957). The procedure is also described in some detail in the present report in Chapter 6.

After dichotomising the response categories for
each item phi-coefficients are calculated.

The items are now plotted in a new two-way table with Thurstone scale values on horizontal axis and the values of phi-coefficients on the vertical axis. The aim is to select most discriminating items (with high phi-coefficients) from each of the Thurstone scale intervals. At this point, the final selection of statements takes place. Edwards and Kilpatrick divided their nine-point Thurstone scale into half scale intervals. Only 7 of these intervals contained statements and they selected four statements with highest phi-coefficients from each of these seven intervals, obtaining a total of 28 statements.

The scale continuum can be divided into larger or fewer intervals. The width of the interval depends upon the distribution of Thurstone scale values, the number of available statements and number of statements desired for the final scale. The investigator should also set a standard for the minimum phi-value acceptable. Edwards and Kilpatrick did not select any item having phi-coefficient below .58.
The above writers further prepared two parallel forms of 14 items each and gave them to a new group of subjects to test for scalability. The subjects were asked to respond in terms of the original six categories. The coefficient of reproducibility for Form A was .875 and for Form B, .872.

Edwards (1957) claims that this method being a synthesis of Thurstone, Likert and Guttman methods possesses certain advantages not present in any of these methods separately. It provides an objective basis for the selection of scalable items which Guttman method alone fails to do. Moreover, this technique helps in eliminating the least discriminating items from the same scale interval, having comparable Q values, which Thurstone method fails to do. Thurstone method also, by inclusion of "neutral" items tends to lower reliability and decrease reproducibility of items. Thus, this method makes use of Thurstone procedure, retains Likert procedure and meets Guttman requirements.

There have not been many investigations to determine the strength and weaknesses of this technique. But,
Remmers (1954) reports that in few applications of this technique so far published, it has lived up to expectations quite well. Passey and Pennington used this technique to develop scales toward prohibition and found it useful.

7. Factor Scales: Factorial methods are used to determine the number of dimensions involved. A large number of attitude items are collected. They are given to a sample of subjects to respond in terms of 5 or more point scale. Inter-correlations among the items are obtained and factor analysed. Each of the major factors constitutes a separate scale. The items which relate most prominently to a factor can be used to construct a scale. This method is not popularly used because of the statistical labours involved.

8. The generalised scales: Remmers and Silance (1934) hold that the Thurstone method though very good is a laborious one. To construct a separate scale for each issue will involve considerable labour. They suggest a modification of the Thurstone technique which retains the primary advantages of his
scaling procedure and, at the same time, enables the measurement of many more attitudes with no increase of labour. This method assumes that attitudes towards any one of a large group or class of objects can validly be measured on a single scale. For example, the above authors developed a scale which can measure attitudes towards any racial group like Americans, Russians, Germans, Chinese, etc. The essential feature is, that the statements should be such which would apply to any member of such a large class of subjects. The procedure of construction is the Thurstone technique.

9. The Social-Distance Scale: One of the landmark in the history of attitude measurement was the scale of social distance devised by Bogardus (1925 A, 1925 B). Bogardus developed this technique for measuring attitudes towards different national groups. His social distance scale was made up of such statements as "I would willingly admit members of (the given race) to, 1. close kinship by marriage; 2. to my club as chums" etc. The statements thus expressed from a close relationship to a distant one or none at all. Bogardus was interested in measuring
degrees to which various racial or national groups were accepted or rejected. Instead of favourable and unfavourable attitudes, he conceived the problem in terms of "degrees of distance". The degrees were not determined in terms of equal units, methods of doing so had not yet been established. Later on, however, Bogardus (1933) took into account the requirement of determining scale units. He asked 100 judges to rate 60 statements, according to the amount of social distance expressed by each statement, on a 7 point scale. Then, he selected items for the final scale. Bogardus type scales apply only to classes of people and it is concerned with action rather than general feelings.

10. Behaviour Scales: Rosander (1937) and Pace (1940) argue in favour of a behaviour scale rather than a simple statement scale. Rosander, (1937) following the Thurstone technique, constructed a behaviour scale. Such a scale attempts to project the individual into a real life-situation. The individual is presented with certain situations and asked what he would do when actually faced with
An example of Rosander scale is: "you are reading in a public library. A negro comes in and sits down beside you:

a. you rise and go to another table;

b. you keep right on with your reading memories;

c. you complain to the librarian;

d. you leave the table at once."

Rosander holds that such a situation scale will give reliabilities between .80 to .90.

Pace (1940) also incorporates the use of stated behaviour rather than stated opinions to test individual attitudes. He compared results obtained by a scale using stated behaviour as the indicators of attitudes and another parallel scale using stated opinion and found the situation scale more discriminating than the opinion scale.

Various methods of scale-constructions are discussed above. Green (1954) mentions two more scaling methods, "the unfolding technique" and "latent structure analysis". But, these are scaling models rather
than scaling methods.

The critical evaluations of each method has already been cited from place to place. Some writers strongly support "Thurstone method", while others support "Likert method" and still others support the other methods. Saffir (1937) and Riker (1944, 1945) emphasize that the choice of method should be governed by matters of conveniences. Bearing in mind the pros and cons of each method, choice should be made taking into account the attitude variable under consideration, the type of scale wanted, nature of the sample and convenience in terms of time and other facilities.

Some Salient Features of Attitude Scale Construction:

When confronted with the task of attitude scale construction, the investigator must bear in mind certain practical problems encountered. Listed below are certain salient features of an attitude scale which must be given due consideration while proceeding to construct a scale.
1. Formulation of Statements: The attitude scaling methods necessitate the use of opinion or statements. All attitude scales comprise of phrases or propositions to which the respondent indicates his agreement or disagreement. The scale must sample all possible shades of opinions or beliefs dealing with several aspects of the issue under consideration. The opinions or beliefs incorporated into the form of statements are highly sensitive to wording. The selection of statements should, therefore, be based upon a very careful scrutiny. Preparation of statements becomes one of the most important aspect of measurement.

The statements for the scale can be selected from books, periodicals, journals, magazines and other allied literature. Previous scales constructed on similar issues are also consulted. Statements may be formulated by a sample of subjects, judges, experts, students or colleagues, etc. The investigator himself formulates and modifies the statements obtained from the above sources.

Wang (1938) lists 16 criteria to be observed in
I. The statement must be debatable;
II. It should belong to the same attitude variable;
III. The statement must be susceptible to only one interpretation;
IV. Avoid double-barreled statements;
V. The statements should be short, rarely more than 15 words;
VI. The statement should be complete, denoting a definite attitude towards a specific issue;
VII. Only one complete thought should be expressed in one statement;
VIII. Avoid grouping sentences as one statement;
IX. Statements should be clear and direct;
X. Avoid or use sparingly such words as "only", "more", "just", as they are likely to be bimodal;
XI. Avoid colourless expressions lacking in effect;
XII. Use simple sentences rather than complex or compound;
XIII. Use complex sentences rather than compound;
XIV. Use active rather than passive voice;
XV. Use the term at issue as the subject of the sentence even if passive voice then becomes necessary;

XVI. Avoid high sounding words or uncommon expressions.

Thurstone (1928) lists the following criteria:

I. Statement should be brief;

II. It should be such that it can be accepted or rejected;

III. The statement should be such that its acceptance or rejection should express a definite attitude of the respondent;

IV. Double-barreled statements should be avoided except as examples of neutrality, if no better neutral items are available;

V. The statements should belong to the attitude variable that is to be measured.

There are many others who emphasize the importance of care to be taken while formulating and editing the
statements.

2. Length of the Scale: The question of the length of a scale is an important and pertinent one. The number of statements in available scales varies from a few to a hundred and more. It is usually determined in arbitrary manner. Ferguson (1952) used 20 to 78 statements. Thurstone and Chaves (1929) scale to measure attitude towards Church consists of 45 statements. Table I lists a few scales constructed and number of statements each scale is comprised of. This varies from scale to scale and, investigator to investigator. Increasing the length of the scale increases reliability but, at the same time it increases inconvenience to the investigator as well as to the respondent.

As to how many statements should be retained in the final scale depends upon the method used in constructing and the number of response categories used. In a Likert type 20 item scale, there are 100 possible responses, whereas in Thurstone type 20 item scale there are only 20 responses.
So, the Likert type scale with fewer items may actually be longer than a Thurstone scale (Ferguson 1952). Selection of statements and making up the final scale should depend upon the following criteria (Krech et al. 1962):

I **Discriminating function:** The items should discriminate between persons holding opposite attitudes;

II **Sharpness of discrimination:** The items should discriminate as sharply as possible;

III **Discrimination along the entire scale:** The items should afford fine discrimination of favourable and unfavourable items of all shades along the entire scale.

IV **Reliability:** There should be minimum number to give maximum reliability.

3. **Number of intervals or number of response categories:** Another important question that is raised is as to how many intervals to use in Thurstone method and how many response categories to use in Likert method. Great variability is found in this aspect as well.
Thurstone and Chave (1929) used 11 categories. Ferguson (1952) has used 9 and 7. Webb (1955) reports those who have used as many as 17 and as few as 5 or even less. Seashore and Havner (1933), Edwards and Kilpatrick (1948), Smith (1932) and Dudycha (1941) used 11 categories. Rosander (1937) used 12 piles. Panda and Kanungo (1962) and Marks (1943) used 5 categories. There are many others who used other variations. It is believed that reliable scale values can be obtained by using fewer subjects and fewer intervals than were first used by Thurstone (Webb 1955).

Number of response categories used in Likert type scales is also exhibited by variations. Likert (1932) recommended 5 response categories. Kerr (1952), Brown and Lowe (1951), Banerjee (1962) and many others used 5 categories. Edwards and Kilpatrick (1948) used 6 response categories. There are others who use only 3. But the usual practice is to use 5 or 6 categories. More than 6
4. The fourth problem faced by attitude researchers is the problem of the size of sample to be used. Thurstone method requires judges and Likert method respondents to serve as the standardisation group.

Thurstone and Chave (1929) used 300 judges to obtain scale values of 130 statements. According to Edwards (1957) a small number of judges can give as reliable scale values. Edwards and Kenney (1946) report a correlation of .35 between the scale values for 129 statements obtained from a group of 72 and 300 judges. Ferguson (1952) used 25, 50, 75, 100, 125, 150 and 200 judges. The writer like Edwards and Kenney concludes that reliable evaluation can be secured by as few as 25 judges. Webb (1955) reports writers who have used 200 to 300 judges as well as those who used 50, 25 or even 2, 5, 10 and 15. There is no hard and fast rule regarding the number of judges to be used. Scale values of high reliability can be obtained on the basis of as
few as 15 to 20 judges ( Wyb 1955 )

The main problem in Likert technique centers around establishing the cutting points for selecting the high and low criterion groups for item validation. Variations are observed in the percentage of locating the cutting points. Ferguson ( 1952 ) reports Likert as having used upper and lower ten percent in his study. Studies show that some research workers take 27 percent ( Hartman 1938 ), some take 25 percent ( Muthayya 1958 ) and some take 10 percent ( Sampson and Smith 1957 ), as high and low criterion groups. There are many other variations around these figures. Here again, there are no hard and fast rules for determining the number of subjects or percentage at which to locate the high and low scorers. Reamers ( 1954 ) recommends top and bottom 10 percent ( or 27 percent ) or any other percent. Ferguson ( 1352 ) speaks of top and bottom 25 percent or any other convenient proportion. Edwards and Kenney ( 1946 ) recommend upper and lower 10 ( or some other ) percent.
Edwards (1957) proposes 25 (or some other) percent. It depends upon the investigator to arbitrarily select the cutting points.

5. Another question that poses before the investigator is about the arrangement of items in the final scale. This problem is faced when the scales are constructed by the method of equal-appearing-intervals, because here the items have specific scale values. Should the items be arranged in a random or serial order of scale values? Dunlap and Kroll (1933) maintain that the arrangement of items, in a scale, whether random or in serial orders of scale values has no effect upon the mean, S.D. or reliability coefficient. Day (1941) holds that arranging the items in Thurstone technique, in serial order would facilitate scoring. It may, however, be realised that in methods like the scale-discrimination technique, where each statement has a specific value but which is not utilised while scoring, it would be better to arrange the items in a random order. This will permit response sets or any possible clues which the subject may get from the serial order of items.
In Likert technique, for example, the arrangement of items in sets of favourable and unfavourable ones would, no doubt, simplify scoring but is also likely to create response sets and vitiate results.

6. Problem of application: A debatable point is whether or not the respondents should be requested to sign their names. Study by Ash and Ambramson (1952) proved that anonymity or identification does not affect attitude scores. The use of signatures depends upon the type of attitude study, type of subjects and the social milieu in which the test is taken (Day 1941). If the test calls for judgment about some highly controversial issues about which there is a prevailing norm in society, it would certainly be better not to request the respondent to give his name. As far as possible anonymity should be censured to give the respondents the liberty to express their views.

7. The question still to be answered is about the value of neutral items in a scale. Why neutral items
items are included in the scale and what purpose do they serve? Edwards (1946) found that the neutral items in Thurstone method tend to have i. high Q values, that is, are ambiguous, ii. tend to be irrelevant, iii. express an attitude of ambivalence or indecision and iv. indifference or apathy. They were also found to be less discriminating, that is, they do not differentiate between the high and low scorers. When Thurstone items are scaled by Likert method a 'U' shape curve is found. This indicates that the Thurstone items with scale values at the two ends are associated with high discriminating powers and the items toward the middle of the continuum, that is, neutral items have low discriminating power. The neutral items are likely to be endorsed by both high and low scorers. As such, they are non-differentiating. It has been demonstrated (Ferguson 1941) (and Edwards 1946) that Likert selected statements tend to be those falling outside the neutral section of Thurstone continuum. Likert also speaks of only favourable and unfavourable items for selecting the initial
The neutral items contribute to errors of reproducibility when included in a Guttman type scale (Edwards 1957). Having the likelihood of being endorsed by the high and low rank-order scorers cutting points for these are difficult to establish. Neutral items have low discriminating power and consequently low reproducibility. In one study, Edwards (1957) found that by the inclusion of neutral items the coefficient of reproducibility could not be calculated but, when the papers were rescored, eliminating neutral items the coefficient of reproducibility was .01.

Some authors hold that a neutral attitude is not possible. (Reuter 1923, Bogardus 1950 and Sargent and Williamson 1958). An attitude is either for or against, positive or negative but not neutral. It may be one of desire or aversion but it is never of indifference. Kundu (1962) also argues that the neutral point in attitude does not exist. All attitudes are amenable to a two fold
division of favourable-unfavourableness. It is like
the border between the two countries and a man
must be either on this side or that side. Statisti-
cally the zero point exists but psychologically
it is not applicable. The attitudes of human beings
can not be indifferent. Through learning, sugges-
tion and imitation one tends to develop pro and con
attitudes.

Whereas some writers deny the existence of neutral
attitudes, others (Kroch et al. 1962) emphasize
the establishment of a neutral region. An individual's
attitude is characterised by pro and conness. That
means the attitude scales have a region where the
sign changes and this is the neutral region. To
one side of this region attitudes grow more positive
and to the other side more negative. The determination
of this region is considered very important in the
measurement of attitudes. Thurstone method ensures
this region by providing the middlesmost or neutral
category. In Likert technique the interpretation
of this point is ambiguous. It is wrong to assume
that the neutral region in a Likert scale corresponds to the middle point of the possible range of scores. The middle point scores can be achieved in two ways - by taking a neutral position on most or all of the statements, or by taking a strongly favourable position on some and strongly unfavourable on other items. Thus, the neutral position cannot be adequately determined. For crude purposes, of course, it may be sufficient to note the persons falling in the middle as fairly close to the neutral region. But this does not mean that they have no attitude. It only shows they are not extremists.

Guttman and Suchman (1947) describe a method to determine the neutral region. A person holding an extreme attitude is more likely to feel more intense and more emotional about it, than the one holding a less extreme attitude. Plotting the scale scores on horizontal axis and the corresponding intensity on the vertical axis gives a 'U' shape curve, intensity being lowest in the middle and highest at the extremes. The lowest part of the curve is the
neutral region. This is the region where the sign changes. To one side lie the persons with pro attitudes and to the other side persons with con attitudes.

Thus, where some authors deny the existence of neutral region, others emphasize it.

In the present investigation though the neutral items are not included as even otherwise they automatically get eliminated when item-analysis procedures are applied. The author has, no doubt, kept a neutral point in the scoring system. It becomes essential to provide the subject with a middle most point which he can chose when he is unable to make any decision of agreement or disagreement with the statement. Sometimes, an individual’s attitudes on a particular issue or issues are not well crystallized. They may be in the stage of formation or transformation. Hence, it becomes difficult for the subject to state his agreement or disagreement with it. Moreover, certain issues are characterised by a state.
of indiscision for most of us. Hence, a convenient neutral point becomes essential in the response categories.

Requirements of an Adequate Attitude Scale:

The problems of reliability and validity are the persistent problems in attitude measurement. The measuring instrument should be reliable as well as valid. These two aspects are the most important requirements of an attitude scale.

Validity: Validity refers to the degree of success with which a test or the instrument measures what it is supposed to measure. Before applying the scale to measure attitude, it should be ensured that the scale is a valid measure of attitudes. It should tap the same attitudes for which it is meant. There are several methods of determining validity:

1. Content Validity: When we proceed to construct a scale we begin with a definition of the attitude
variable and a sample of items related to that attitude variable. To be valid, the scale must live up to our definition and the scale items should give a fairly representative sample of beliefs, opinions and action tendencies pertaining to the attitude object under consideration. This is called intrinsic or constant validity. This type of validity may be evaluated by a group of experts as is done in Thurstone's method of equal-appearing-intervals. The judges give their ratings as to what shade of attitude the item measures.

2. Method of interval consistency: Likert method provides for validation of scales by this procedure. Certain checks of interval consistency are employed to test validity, for example, correlating each item with the total score. High correlations show that each item is measuring what the scale is measuring. The items hang together and all of them measure the same attitude variable. The discriminating power of an item also argues for its validity. When an item differentiates sufficiently between the two extremes, high and low groups, it is said to
measure the particular trait or attitude which makes for these differences. In this method every item is validated which ensures the validity of the total scale.

3. "Known group" validation: If we can differentiate certain groups whose attitudes are known to us, for instance, if we know that a particular group is highly religious and another group less religious, then our scale to measure religious attitudes would be considered valid if the highly religious group members get high scores and low religious group members get low scores. In the internal consistency between procedure of validation every individual item is tested for its discriminating power, that is, its ability to discriminate between the high and low scorers on the total scale. But, in the "known group" method of validation the entire scale is taken together, that is, the total score should be able to discriminate between the high and low groups. Moreover, these two groups should be known to be 'high' and 'low' by some other criterion and then the test is applied to find out whether it
validly separate the two groups. In the method of internal consistency, the groups are designated 'high' and 'low' by the scale score and then each item score is tested for its consistency with the scale-score.

4. Correlation with external criteria: The scale should correlate highly with some other criterion already available as an indicator of the attitude. The external criterion may be of the following type:

i. Correlation with other scales: If some other scales are available to measure the same attitude, then the present scale, which might have been developed by some other method, should correlate highly with the available scales.

ii. Agreement with ratings based on personal knowledge: The attitude score of the individual obtained by the scale should be in agreement with the ratings obtained of the subject on the attitude object by friends, acquaintances, teachers, etc, who know him personally.

iii. Agreement with self-ratings: The attitude score as indicated by the scale should be in agreement
with the individual's own ratings of his stand on the attitude variable.

It may, however, be noted that while taking ratings as an external criterion one must be very cautious, as ratings may be biased due to various reasons.

5. Agreement between the attitude score and observation of the individual's day-to-day behaviour: For example, to find out whether the religionism score obtained by an individual is in agreement with his day-to-day behaviour, it may be noted whether he visits the place of worship regularly, whether his daily routine is characterised by daily prayers and observation of religious rituals.

It must again be borne in mind that an individual's observed behaviour may not always be an indicator of his true attitudes. The question of discrepancy between attitude and action has not been resolved.

6. Factorial validity: Factorial method has also been used to determine the validity of the scale. This method consists of correlating the scale with the
factor making up the scale. Ferguson (1952) used this method to establish the validity of his three scales arrived at by factor-analysis and item-analysis methods.

Any of the above methods or a combination of some can be used to determine the validity of the scale. Table 1 shows the various procedures used by investigators to validate the scales. It is obvious from the table that some writers use the "known group" method (Cristle 1940, Levinson and Sanford 1944, etc.) while others use the method of "internal consistency" (Benerjee 1960), correlations with external criteria like other scales or ratings (Muthyra 1966, Lentz 1935, etc.) and correlations with actual behaviour (Thurstone and Chave 1929, Brown and Lowe 1951) have also been used. Some writers also take opinions of judges to validate the scales (Menesas 1958, Ledbetter and others 1949) and factorial validity method has also been used (Ferguson 1952). A combination of a few of the above methods is also not uncommon (Thurstone and Chave 1929, Sampson and Smith 1957, Brown and Lowe 1951, Ferguson 1952 etc.).
Reliability: Reliability is generally considered as the consistency of repeated measurements of the same events by the same process. It is an index of the extent to which repeated measurements yield similar results. A valid test may not always be reliable. A yardstick made of highly elastic rubber measures lengths of objects and is, therefore, valid but not reliable. Repeated measurements may give different results depending upon how much you stretch it. The methods commonly used to determine the reliability of test scores are as follows:

1. Test-retest method: Repetition of the test is the simplest method of determining reliability. The same individuals respond to the test twice and the two measurements are correlated. Such a coefficient of reliability is known as the coefficient of 'stability'.

This method has its limitations too. Errors may creep in due to the changes in conditions either in
the examiner or examinee or environment during the two administrations. Skill of knowledge acquired during the first administration or the memory of the first test may also vitiate the results.

2. Alternate form or parallel form method: This method consists of giving two parallel forms of the same test and determining the coefficient of correlation between the two sets of scores. This is often known as the "correlation of equivalence". The difficulty in this method is that error variance in this method may be due to the changes in content of the two tests. Moreover, the knowledge and skill acquired in the first test may be transferred to the second test and vitiate the results.

3. The split-half method: In this method the test is broken in two equivalent parts and the correlation between these half-tests is computed. This is often called the correlation of "internal consistency". The splitting may be based upon random (odd-even numbered items), on first half versus second half or matched halves techniques in which the two halves are matched with respect to standard
diviation, mean etc. In split-half, it becomes essential that the two halves resemble each other and each should faithfully represent the total test. This method is used when it is not feasible to construct two parallel tests or repeat the same test twice.

Split-half method is regarded by many as the best of the methods for measuring test reliability. It is easier and quicker as only one test is needed and scores are also obtained in a single administration. A marked disadvantage of this technique is that chance errors may affect the both halves in the same manner and thus make the reliability coefficient high. It is also criticised as giving overestimate of test reliability.

The three approaches outlined above have the common goal of deriving two sets of scores for the test on the same sample for computing coefficient of reliability. In split-half method Spearman-Brown formula is usually applied to estimate the reliability.
of the whole test.

4. Kuder-Richardson formula for estimating test reliability: Dissatisfied with the split-half method, Kuder and Richardson proposed a new procedure based on item statistics (Guilford 1954). This is also known as the method of "rational equivalence".

This method splits the test into $n$ parts of one item each. By split-half method the test can be split into two in a number of ways. If we take a mean of all possible splits this will give K-R formula. Kuder-Richardson gives number of formulas but the K-R formula 20 is most commonly used. Pressel and Gerguson give modifications of this formula. Guilford (1954) has discussed these methods in detail. The K-R method usually gives a lower-bound estimate of reliability. According to Garrett (1960) K-R method though superior to split-half, the actual difference between the reliability coefficients found by the two methods is small and negligible.

The choice of method depends upon the nature of the
test, the convenience and use of test. Table 1 shows that the first three methods"parallel forms, test-retest, and split-half"are usually used in attitude tests. The range of reliability coefficients obtained is also shown vide Table 1.

BRIEF REVIEW OF PREVIOUS ATTITUDE SCALES:

Ever since the scaling methods came into prominence there have been various efforts to devise scales to measure the specific attitude variables. The publication of the method of equal-appearing-intervals in 1929 by Thurstone and Chave resulted in a series of scales constructed by Thurstone and his associates and others to measure attitudes towards the various social issues. Even before the scaling methods came into light, there had been efforts to measure attitudes. Zeleny (1926), for example, prepared an 'opinion test' consisting of 34 items in 1926. He scored his test in a different manner called the "forward and reverse manner". Zeleny also made use of judges to evaluate statements but not in the sense in which Thurstone did. Bogardus (1925 A, 1925 B)
also devised a test to measure social distance, before the use of scaling methods was introduced in the field.

Thurstone and his associates have constructed a number of scales to measure attitudes toward war, prohibition, God, Church, capital punishment, birth-control, communism, patriotism, movies, etc. There are many other scales on a wide variety of social issues which it is not possible to list all here. Table 1, however, gives a brief review of a few social attitude scales constructed by various writers.

The methods used in construction and reliability-validity determination are also cited alongside.

The Table shows only a few of the numerous scales so far constructed. It is evident that efforts in our country in this direction are very meagre.
<table>
<thead>
<tr>
<th>Author</th>
<th>Scale</th>
<th>Method</th>
<th>Number of items</th>
<th>Validity</th>
<th>Method</th>
<th>Co-efficient</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thurstone and Chave (1929)</td>
<td>Scale to measure attitude toward Church</td>
<td>Equal-appearing-intervals</td>
<td>45</td>
<td>Known groups and self ratings with actual behaviour</td>
<td>Parallel</td>
<td>.92</td>
<td>Thustone and his colleagues after the publication of this scale constructed a series of scales on issues like War, Birth-Control, Evolution, God, Communism, Capital Punishment, etc.</td>
</tr>
<tr>
<td>Seale, E. S. (1933)</td>
<td>Scale to measure Social distance</td>
<td>Judges</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zeleny (1926)</td>
<td>Opinion test (Two forms)</td>
<td>Empirical method of point response</td>
<td>34</td>
<td>Consistency of in each answering in forward and reverse manner</td>
<td></td>
<td>.81 to .89</td>
<td></td>
</tr>
<tr>
<td>Griswol, M (1940)</td>
<td>M-S Scale</td>
<td>Empirical method of point response</td>
<td>65</td>
<td>Known groups</td>
<td>Split-half</td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td>Author/Year</td>
<td>Scale/Method</td>
<td>Sample Size</td>
<td>Split-Half</td>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------</td>
<td>-------------</td>
<td>------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day &amp; Quickenbush (1942)</td>
<td>War scale (generalised scale) Equal-spacing-intervals</td>
<td>13</td>
<td>Split-half .90 to Used 50 Judges to sort the items in 9 piles</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levinson and Sanford (1944)</td>
<td>Anti-Semitism</td>
<td>52</td>
<td>Known group</td>
<td>Split-half .96 The scale consisted of only negative items. Method of internal consistency like that of Likert was used. D· values were calculated but a different scoring system was adopted</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conrad and Sanford (1943)</td>
<td>War Optimism (two scales)</td>
<td></td>
<td>10 items in scale</td>
<td>1. split-half .40 to .70 2. alternate forms</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kilpatrick (1936)</td>
<td>Feminism Scale (3 forms)</td>
<td>sorting by Judges</td>
<td>00 for Known groups</td>
<td>1. test-retest .94 categories of feminism and anti-feminism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>Measure of Attitudes</td>
<td>Method</td>
<td>Items</td>
<td>Subgroups</td>
<td>Reliability</td>
<td>Notes</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------</td>
<td>-------</td>
<td>------------</td>
<td>-------------</td>
<td>--------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Sampson and Smith (1957)</td>
<td>to measure attitudes towards world-mindedness</td>
<td>Likert Method</td>
<td>30</td>
<td>Known groups</td>
<td>.95</td>
<td>The author used 11 piles parallel-forms</td>
<td></td>
</tr>
<tr>
<td>Stagner (1942)</td>
<td>War scale two forms</td>
<td>Method of absolute judgement</td>
<td>27</td>
<td>in the lower form</td>
<td>.95 to .96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dudyacha (1943)</td>
<td>to measure attitudes towards dependability</td>
<td>equal-appearing intervals</td>
<td>27</td>
<td></td>
<td>.95</td>
<td>The author used 11 point continuum</td>
<td></td>
</tr>
<tr>
<td>Williamson and Darley (1937)</td>
<td>1. social preference test 2. socioreaction test</td>
<td>A combination of Thurstone &amp; Likert methods</td>
<td>40</td>
<td>Known groups</td>
<td>.92</td>
<td>The author used 5 response categories.</td>
<td></td>
</tr>
<tr>
<td>Smith (1932)</td>
<td>to measure attitudes towards prohibition</td>
<td>equal-appearing-intervals</td>
<td>45</td>
<td>Known groups</td>
<td>Parallel-forms</td>
<td>The author used 11 piles continuous</td>
<td></td>
</tr>
<tr>
<td>Kerr (1952)</td>
<td>Liberalism-conservatism in 5 areas (5 scales)</td>
<td>Likert method</td>
<td>76</td>
<td>Known groups</td>
<td>Split-half</td>
<td>Used 5 point response</td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** 1. The author used 10 percent test-retest as the criterion groups. 2. The author used 11 piles parallel-forms.
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown &amp; Lane (1951)</td>
<td>Inventory of religious beliefs</td>
<td>Likert Method</td>
<td>15</td>
<td>1. opinions of experts 2. Known groups 3. Correlation with personal data as indicative of behaviour</td>
<td>.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senefee (1938)</td>
<td>to measure radicalism-conservatism attitudes</td>
<td>Opinions of Judges</td>
<td>40</td>
<td>opinions of Judges</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levinson (1946)</td>
<td>Political-economic conservatism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thurstone (1959)</td>
<td>to measure attitudes toward movies</td>
<td>equal-appearing intervals</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferguson 1939, 1944, 1952</td>
<td>Tri-E scales 1. Religionism 2. Humanitarianism 3. Nationalism (2 forms of each)</td>
<td>factor analysis and item analysis</td>
<td></td>
<td>1. Factorial method Alternative Forms .88 to .95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferguson</td>
<td>1. Religionism 2. Humanitarianism 3. Nationalism (2 forms of each)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pace (1940)</td>
<td>S-R survey and opinion survey to measure radicalism-conservatism</td>
<td></td>
<td></td>
<td>odd-even S-R survey with the opinion survey .824</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rosendor (1937)</td>
<td>Behaviour scale to measure attitude toward Negroes (two forms)</td>
<td>Equal-appearing intervals</td>
<td>22</td>
<td>1. Known groups 2. Correlation with another scale Parallel Forms .98</td>
<td>The author took 50 judges to sort and 11 piles. Instead of statements he used life situations and reactions to them.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Factor analysis of intercorrelation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portfield (1938)</td>
<td>to measure attitudes towards war</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marks (1943)</td>
<td>Two scales for measuring attitudes toward Whites and Negroes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32 in each scale,</td>
<td></td>
</tr>
<tr>
<td>Vetter (1930)</td>
<td>to measure social and political attitudes (36 sections in the scale)</td>
<td></td>
<td></td>
<td>5 in each section</td>
<td></td>
<td>160 items in total</td>
<td></td>
</tr>
<tr>
<td>Lenz (1935)</td>
<td>C-R opinionnaire (two forms)</td>
<td></td>
<td>1. Interval consistency 60 in each Judges</td>
<td></td>
<td>1. opinions of 50 in each form 2. correlation with self-ratings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eysenck (1954)</td>
<td>S and T Scales</td>
<td></td>
<td>Factor analysis of item intercorrelation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The items were not satisfactorily weighted.

He gave the judges 5 point rating scales.

The method of construction of C-R opinionnaire is described by Nelson (1939-III).
<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ledbetter and others</strong>&lt;br&gt;(1949)</td>
<td>Opinionnaire on Democracy (two forms)</td>
<td>65 in each form</td>
<td>Judgments of experts</td>
<td>Split-half .93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Passey and Pennington</strong></td>
<td>5 scales on Alcoholism (2 forms each)</td>
<td>Scale discrimination technique</td>
<td>12 in each form</td>
<td>Parallel .91 to .96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Banerjee (1962)</strong></td>
<td>4 scales A, C, A₂, H, A₃, R, A₄, S</td>
<td>Likert method</td>
<td>40 to 50</td>
<td>1. 'Jury' validity (opinions of judges)</td>
<td>Split-half .88 to .96</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Kuthy (1958)</strong></td>
<td>TFL scale (traditional Family Ideology) (Based upon Levinson and Huffman's Traditional family ideology scale)</td>
<td>Likert method</td>
<td>40</td>
<td>Correlation with external criterion (friends' ratings and a social distance scale)</td>
<td>Split-half .912</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pandy and Kamungo (1962)</strong></td>
<td>To measure attitude toward motion pictures appearing-intervals</td>
<td>30</td>
<td>Split-half .84</td>
<td>He used 5 piles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>To measure religious attitudes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>--------------------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Adinarayan and Rajamanickam (1962)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>Das Gupta and Ghosh (1960)</td>
<td>To measure attitude toward social work, Thurstone and Likert (both independently)</td>
<td></td>
<td></td>
<td>19</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sinha and Pandya (1960)</td>
<td>9 attitude variables</td>
<td></td>
<td></td>
<td>1 item for each variable</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Banerjee and Kundu (1960)</td>
<td>To measure teachers' Likert type attitudes toward authority</td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dhawan and Shrivastava (1959)</td>
<td>I-R Scale (Idealism-</td>
<td></td>
<td></td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>scialism)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The construction of attitude scales is analogous to determining benchmarks on any physical scale. A typical attitude scale consists of a series of statements on the aspect under study. Some of the principal methods of scale construction are elaborated in this Chapter. A critical evaluation of each is also presented alongside. This will give the reader an insight into the magnitude and importance of the area of attitude.

The choice of the method to be used for constructing social attitude scales is determined with reference to the specific problem in hand. No doubt, the governing factors here are the attitude variable, sample available, and convenience of the investigator. An attempt has also been made to stress some of the salient features of attitude scale construction. These are indispensable aids for successful work in this area. Finally, a tabular analysis of the relevant previous studies has been presented. This will help in the evaluation of the present study in the light of these.