CHAPTER IV

DEVELOPMENT OF TEACHER RATING SCALES
CHAPTER IV

DEVELOPMENT OF TEACHER RATING SCALES

4.1 Rating Scale and its procedure of construction:

A.S. Barr et al. (1958) state that Rating is a term applied to expression of opinion or judgement regarding some situation, object or character. Opinions are usually expressed on a scale of values. Rating techniques are devices by which such judgements may be quantified.

According to Thorndike and Hagen (1955), rating procedure attempts to get appraisal on a common set of attributes and traits for all raters and ratees and to have these expressed on a common quantitative scale. The most common pattern of rating presents the ratee with a set of trait names, range of numbers, adjectives or descriptions that are to represent levels of degrees of possession of the traits assigning him to judge these to fit best.

The rating scale is a very useful device in assessing quality - specially when quality is difficult to measure objectively. For example, 'How good was his teaching?' is a question which can hardly be answered objectively.

Rating scales record judgements or opinions and indicate the degree or amount. Descriptions of different degrees of quality are arranged along a line from high to low.
or otherwise, and this line is the scale.

This most commonly used instrument for making appraisals has a large variety of forms and uses. Typically, they direct attention to a number of aspects or traits of the thing to be rated, and provide a scale for assigning values to each of the aspects selected. They try to measure the nature or degree of certain aspects or characteristics of a person or phenomenon through the use of a series of numbers, qualitative terms, named attributes or verbal descriptions.

Rating scales have been successfully utilized for the following purposes:

I. Teacher Rating - for selection, evaluation and prediction.
II. Personality Rating - for various purposes.
III. Criterion Rating - for testing the validity of many objective instruments like paper-pencil inventories of personality.
IV. School appraisal - including appraisal of courses, practices and programmes.

Making proper use of rating scales is very flexible and simple procedure but it depends on the judges instead
of independent criteria and therefore, the use of rating scales is not wholly objective. However, some of the important points, about the construction and use of Rating Scales, are given here for the research workers to be borne in mind before they start working on them. They are as follows:

1. Rating scales include three factors: (i) The subjects or the phenomena to be rated, (ii) the continuum along which they will be rated, and (iii) the judges who will do the rating. All these three factors must be selected very carefully.

2. The subjects or phenomena to be rated are usually a limited number of aspects of a thing, or of traits of a person. Only the most significant aspects for the purpose of the study should be chosen.

3. (a) The usual way to get qualitative-cum-quantitative judgement on the selected aspects of a thing or person is to set up five to seven categories, along the continuum either from superior to inferior (through above average, average and fair) or from excellent to poor (through good, average and below) or always to never (through frequently, occasionally and rarely).

   (b) Another procedure, however, establishes position in terms of specific behavioural or situational
descriptions. For example, instead of deciding whether the individual's professional qualities are superior, above average, average, fair or inferior, it may be easier to decide between: 'His teaching method is always interesting' and 'Sometimes his teaching method is very interesting'.

The use of the latter form may be considered with advantage, that is why, the investigator used this form of rating scale in the construction of teacher Rating scales for the present piece of research with a modification in choice of four alternatives instead of three or five.

4. The different degrees of quality are usually adjectives or descriptions. Their meaning should be clearly different from each other, and in going 'up' the scale one should feel that the next description represents better quality than the last. In writing and arranging these descriptions it is good to have the judgement and agreement of several people.

5. Items may be arranged in ascending or descending order from left to right. Ordinarily an odd number of steps should be used and the average or usual quality should be kept in the middle but the researcher did not do this way in this study.

6. At least three divisions of quality must be kept,
because a two division rating scale (as yes/No or bad/good) forms a check-list but not a rating scale. A rating scale may have as many divisions as can be readily distinguished by the raters. Practically, most scales have no more than seven divisions. Usually they contain five divisions. By numbering each division in sequence the descriptions can be converted into arithmetic values for averaging.

7. The rating scale is always composed of two parts: (i) an instruction which names the subject and defines the continuum and (ii) a scale which defines the points to be used in rating.

8. Any one can serve as a rater where non-technical opinions, likes and dislikes and matters of easy observation are to be rated. But only well-informed and experienced persons should be selected for rating where technical competence is required. So an educational investigator should employ only those persons as raters who are either experts in the field or who form a sample of the population in which the scale will subsequently be applied.

9. Pooled judgements increase the reliability of any rating scale. Employ several judges, depending upon the rating situation, to obtain desirable reliability. Individual ratings when combined into a final rating give a safer assessment. For this, the research worker should supply a separate copy of the scale to each rater to mark his
judgment on, independently, to avoid the possibility of contamination of judgment otherwise. The marks should be tallied on a single copy.

10. A single rating index, the average score, can be calculated by adding all the numbers of various raters and dividing the total by the number of raters.

11. One rating index can be compared with the results of other ratings using the same scale if it serves some logical purpose.

4.1.1 Limitations of Rating Scales:

The following difficulties in constructing and using Rating Scales should be noted carefully:

1. It is difficult to convey to the rater just what quality the researcher wishes him to evaluate. An adjective or adverb may have no universal meaning. Brief behavioural statements or situational descriptions which have clear objective connotations should be used in place of vague terms.

2. It is difficult for raters to get rid of the halo effect which causes a rater to carry qualitative judgment from one aspect to other. For example, one tends to rate a person, with a pleasing
personality, high on traits like intelligence or professional interest too. Halo effect appears frequently when the rater has to rate a number of factors on some of which he has no evidence for judgement. It is advisable to keep at a minimum the number of essential characteristics to be rated. Otherwise, it should be provided to the raters to omit rating characteristics that they have had no opportunity to observe.

3. Raters tend to be generally generous. It has been verified that 60 per cent to 80 per cent of an unselected group tend to receive above-average ratings in all traits. The raters are inclined to be unduly generous in rating aspects which they have had no opportunity to observe.

4. Although pooled judgements of many raters reduce subjectivity of individual judgements, yet the average of scores thus arrived at, are by no means either totally objective or highly valid. In interpreting the rating indices, due allowance has to be made of this factor.

4.1.2 Reliability of Ratings

Symonds (1956) concluded that the correlation between the ratings given by two independent raters
for the conventional type of rating scale is about .55. It is possible to pool the ratings by Spearman-Brown Formula. Following estimates are for the reliability of pooled ratings:

- 2 raters correlation: .71
- 3 raters correlation: .79
- 4 raters correlation: .86 (Hagen)
- 6 raters correlation: .92

### 4.1.3 Validity of ratings

Rater biases and rater unreliability operates to lower validity. However, it is usually very difficult to make any statistical test of the validity of ratings. When ratings are being studied as predictors, statistical data can be obtained as to the accuracy with which they do predict.

The validity of a test, or of any measuring instrument, depends on the fidelity with which it measures what it purports to measure. A home made yardstick is entirely valid when measurements made by it are accurate in terms of a standard measuring rod. Similarly, a test is valid when the performance which it measures correspond to the same performance as otherwise independently measured or objectively defined.
The validity of a test is found by comparing the data obtained from the instrument with standard measures. But since independent standards, that is, criteria, are hard to get in mental measurement, the validity of a mental test can never be estimated as accurately as the validity of a physical instrument.

Validity is a relative term. A test is valid for a particular purpose or in a particular situation— it is not generally valid. After World War I, several business concerns used the Alpha Psychological test in the selection of applicants for routine clerical jobs. Those chosen often proved to be poor workers, indicating that the test was not a valid measure of the skills needed in many office jobs.

4.2 Teacher Rating Scales developed by the Researcher

Careful preparation of rating scale is of great value as stated by Cronbach (1943). Three-point and five-point scales obtain more discrimination than the "yes-no" check list. In these types of rating scales, the neutral point is always in between. Generally, the subjects, who want to conceal their real attitudes for one reason or the other known to them only, mark tick on the neutral value of most of the items describing such qualities of those subjects as they do not want to reveal against some social taboos or moral principles.
The investigator developed neither a three-point nor a five-point scale as he did not want to put the neutral point in between. To overcome this difficulty, he used four-point scales arranging all the four choices in descending order of value and leaving no scope for the neutral point. He obtained the ratings of the heads, colleagues and pupils, as raters about the ratees (teachers) on teaching success on two different four-point scales. One he constructed for the use of heads or colleagues and the other for the use of pupils. Generally ratings showed conformity with the variable, that is, teaching success, on which the ratings were obtained.

At the very outset of the study, a well-designed plan, for constructing suitable rating scales for the use of heads, colleagues and pupils to assess teaching success of the teachers, was drawn by the investigator after having consulted relevant literature according to the instructions of the methods of 'rating scale construction' and the long discussions with the experts. These two Teacher Rating Scales - one for the heads or colleagues and the other for the pupils - contained seventy statements in all about seven different sections, each section containing 10 items and each item was further supplied with four choices arranged in descending order.
of value. These scales covered different characteristics of teachers such as teaching methods and knowledge, relationship with pupils and discipline, personality traits, relationship with boss and colleagues, relationship with parents of pupils, attitude towards teaching profession, and hobbies and other interests. Weightage to items was assigned in descending order of value in accordance with their importance in teaching making it a 3-2-1-0 scale. The items of these two-scales were not the same, rather they were different. The items of the Teacher Rating scale for the use of pupils were shorter in size and easier in meaning than those used in the Teacher Rating Scale meant for the heads and the colleagues because the students of XI, X and IX classes were not so much qualified to understand and to interpret longer sentences and to understand difficult words in English though they were explained each and every thing about the rating scale and its procedure of rating, through the instructions written on the first page of the scale and orally by the investigator before use.

Each teacher was rated by the head (Headmaster or Principal of the school), four colleagues (teaching common classes) and five students (of his teaching subject) from the class he was teaching, on these two rating scales. Those teachers rated good and satisfactory and who got total pooled score of these different
ratings higher than the mean score of the total sample were categorized as successful teachers and the others, unsuccessful in teaching profession.

Prior to their final use, the Teacher Rating Scale for the use of heads and colleagues was given to 10 heads of schools and ten teachers and the second Teacher Rating Scale was given to ten students of two high and higher schools, for a try-out. They were asked what else they wanted to say about the teachers they rated in support of their judgement on their (teachers') teaching effectiveness. As a result of their suggestions some alterations in language were made, a number of items were deleted and a few more items were added to the rating scales.

4.2.1 Selection of Statements for the Scale

Current and relevant literature was searched for suitable brief statements that might serve the purposes of the Teacher Rating Scales - one for the heads and the colleagues of the teachers, and the other for their students - in order to assess their (teachers') teaching success. By editing such material, the investigator prepared a list of seventy statements expressive of teacher characteristics under seven different headings. Further each statement was supplied with four alternatives arranged in descending order of value and covering as
far as possible all gradations from one end of the scale to the other.

Special attention to the neutral statements was necessarily given because neutral point was not put in between as the scales used were four-point scales, though the neutral sense was there in the third choice to avoid the danger of breaking the scale in two parts because many times a random compilation of statements fail to produce neutral statements. Thus the whole range of teaching success was fairly covered through different teacher characteristics represented by statements and sections, and gradations represented by four alternatives to each statement.

To avoid the generosity error and halo effect and to have objective results, the rating scales were given to more than three colleagues. The colleagues selected for ratings were those who taught with as well as knew the teachers at least for one year.

To bring refinement in assessing the teaching success of the teachers, the investigator explained its meanings and concept clearly to the heads, colleagues and the students. It was necessary to get reliable and valid opinions of all of them. They were asked to tick one of the four choices of each item and requested to be frank and sincere in their judgement.
In making the initial list of statements suitable for the scales under discussion, several practical criteria were applied in the first editing work. Some of the important criteria are mentioned here:-

I. The statements selected for the rating scales were as brief as possible so as not to fatigue the raters who were asked to tick all the statements leaving no statement of the rating scale unanswered.

II. The four alternatives to each statement were such that they could be endorsed or rejected in accordance with their agreement or disagreement with the understanding of the rater. All the raters - heads, colleagues or pupils - were requested to tick one of the four alternatives of each statement.

III. Every statement was such that it did indicate something regarding the teacher's success in the profession.

IV. Double barreled statements were avoided because they tend to have a high ambiguity.

V. The investigator ensured that at least a fair majority of the statements really belonged to the teaching success variable that is to be measured.
4.2.2 Number and Selection of Judges:

Any number of judges can be selected for the initial try-out of the rating scale but they should be from the field of the variable selected for the study or they should be the persons linked with the subjects in one way or the other so that their judgements may be reliable and valid.

Taking these points in view, the investigator has selected randomly ten heads of schools, ten school teachers and ten students as judges for the initial try-out of the Teacher Rating Scales designed to assess 'teaching success' of the teachers working in high and Higher secondary schools of the Union Territory of Chandigarh.

4.2.3 Item Analysis and Final Rating-Scales

By careful inspection of the statements of the rating scales given to judges for try-out and as a result of their suggestions, alterations in the language of some of the statements were made, alternatives to some of the statements were amended, and a few items were deleted and replaced by others more suitable than the previous ones.

After making necessary changes, amendments and replacements, the investigator was left with the final
scales comprising seventy statements each and having four alternatives to each statement. All these seventy items were divided into seven sections of ten items each. Section A contains ten items pertaining to teaching method and subject matter; section B, to relationship with pupils and discipline, Section C, to personality characteristics; Section D, to the relationship with the boss and colleagues; section E to the relationships with the parents of pupils; Section F, to the attitude of teachers towards teaching profession, and Section G is indicative of the hobbies and other interests of the teachers.

The final scales thus prepared were administered to the heads of schools and colleagues of the teachers concerned as well as to the students, whom they taught in XI, X and IX classes only, to collect data for the thesis.

The investigator faced some difficulties in preparing and administering the rating scales but even then they are the best devices for the purpose of assessing teaching success in such like piece of research. They are widely used these days by most of the educationists and psychologists in the prosecution of their studies.
4.2.4 Scoring of the Teacher Rating Scales:

As these scales contained four choice items, so the four points on the scales were given weighted scores as:

(a) Highest sense  ...  3
(b) Higher sense   ...  2
(c) High sense     ...  1
(d) Low Sense      ...  0

All the weighted scores for each scale rated by the Headmaster/Principal, four colleagues and five students were added together and then divided by ten to get the combined mean score as the teaching success score of the teachers for further analysis.

4.2.5 Reliability of Teacher Rating Scale scores

A test score is called reliable when we have reasons for believing the score to be stable and trustworthy. Stability and trustworthiness depend upon the degree to which the score is an index of "true ability" - is free of chance error.

For example, the Stanford Binet I.Q. is known to be a dependable measure. Hence if a Child's I.Q. is reported to be 110 by a competent examiner, we feel confident that this score is a good estimate of the child's ability to handle tasks like those represented by the test. Scores achieved on un-reliable tests are
neither stable nor trustworthy. In fact, a comparison of scores made of repetition of an unreliable test, or upon two parallel forms of the same test, will reveal many discrepancies—some large and some small—in the two scores made by each individual in the group. The correlation of the test with itself—computed in several ways—is called the reliability coefficient of the test.

There are four methods in common use for computing the reliability coefficient—called the self correlation—of a test. These are:

I. Test-retest (repetition method)
II. Alternate or parallel forms.
III. Split-half technique.
IV. Rational equivalence.

All these methods furnish estimates of the reproducibility of test scores; sometimes one method and sometimes another will provide the better measure.

The investigator employed test-retest method to find the reliability of the Teacher-Rating Scales. He gave the Teacher Rating Scale, for heads and colleagues, to five teachers of one school to get their ratings about a teacher working with them in the same school. Similarly, he gave another Teacher Rating Scale to five
students of X Class to get their ratings about one of their teachers. These tests were quantified with the procedures already explained and the scores were noted down carefully. After fifteen days the procedures were repeated with the same group of raters and the same ratees. This time too their individual quantitative scores were recorded. Then correlations between the first test scores and the second test scores of both the rating scales were calculated and found to be as follows:-

Five colleagues as raters correlation = .81
Five(pupils) as raters correlation = .83

These values of the coefficients of correlation show that the scores of the Teacher Rating Scales for the heads and colleagues as well as for the pupils are reliable. Therefore, the Teacher Rating Scales designed by the investigator himself, for the use of data collection, have proved to be reliable measures of research for this study.

4.2.6 Validity of the Teacher Rating Scales

The validity of any research tool depends upon the trustworthiness with which it measures what it is supposed to measure. A time-piece is a valid measure in telling time if it tells the same time as is indicated
by another clock telling standard time. Validity is a relative term. A test is valid for a particular purpose or in a particular situation - it is not generally valid.

The validity is of many types as content validity, face validity and factorial validity. Content validity is employed in the selection of items in educational achievement tests, and in many trade tests. A test is said to have face validity when it appears to measure whatever the author had in mind, namely, what he thought he was measuring. Rating scales for various traits, attitude scales and even intelligence tests often can claim little more than face validity. Judgements of face validity are very useful in helping a researcher decide whether his test items are relevant to some specific situation or to specialized occupational experiences.

The validity of a test is determined by finding the correlation between the test and some independent criterion. A criterion may be an objective measure of performance, or a qualitative measure such as a judgement of the character or excellence of work done. Intelligence tests were first validated against school grades, ratings for aptitude by teachers, and other indices of ability. One of the best ways is to check test predictions against
actual outcomes. A high correlation between a test and a criterion is evidence of validity provided (1) the criterion was set up independently and (2) both the test and the criterion are reliable.

The index of reliability is sometimes taken as a measure of validity. The correlation co-efficient gives the relationship between the obtained scores and their theoretical true counter-parts. If the reliability co-efficient of a test is .81, for example, \( r_{1c} \approx \sqrt{.81} \) or .90. This means that the test measures true ability to the extent expressed by \( r \) of .90. But here in the case of Teacher Rating Scale for the use of heads and colleagues of the teachers, the coefficient of correlation that is, index of reliability is exactly .81 which shows that this Teacher Rating Scale (for the use of heads/colleagues is valid upto the extent expressed by \( r \) of .90. Similarly, the index of reliability for the Teacher Rating Scale for the use of pupils is .83 which shows that Teacher Rating Scale for pupils is also a valid test to the extent expressed by \( r \) of .91. Lastly, the investigator came to the conclusion that both the rating scales constructed by him have been proved to be reliable and valid measures for the assessment of success in teaching profession.

Though, there are numerous scales, the investigator
prepared these new Teacher Rating Scales for the purpose of data collection for this study. No doubt he followed some hints from other scales but his work is original.