Chapter 4

TOOLS FOR TEACHING EVALUATION; DEVELOPMENT AND CONSTRUCTION

I

Mathematics Teaching Competence Scale (MTCS)

This section examines the various teaching activities that were developed at different parts of the world. Based upon mathematics teaching principles and grade and developmental stages of students involved in our study, a few important activities (21) are conceptualised and MTCS is developed which is an adaptation of BGTCS.

4.1. Classroom Teaching Activities (CTA)

At present effective teaching has been analysed into a few limited well defined components that can objectively be measured. Stanford University has been the pioneer in deriving 14 teaching skills as follows:

1. Passi, B.K., loc.cit., pp. 12-14

(1) Stimulus variation, (2) Set Induction,
(3) Closure, (4) Teacher silence and nonverbal cues,
(5) Reinforcing pupil participation,
(6) Fluency in questioning,
(7) Probing questions, (8) Use of higher order questions,
(9) Divergent questions, (10) Recognising and attending behaviour,

California far west laboratory has derived more skills in addition to 14 skills stated above. These are - "providing feedback", "control of participation", "student initiated questions" and "precuing". Flanders has also derived a few basic skills from a model of "speaking and listening". Case (Baroda, India) has conceptualised 21 teaching skills mostly based on skills developed at Stanford University and California F.W. Laboratory.¹

Criticism

Spelman and Brooks, Travers and many others² have strongly criticised such type of analytical approach to measure teaching effectiveness. Pinsent³ says that teaching effectiveness depends on teacher personality, learner, subject material, specific condition and objectives. Teaching may, therefore, be treated as a "transactional process" depending upon the functions of several uncontrolled variables.

However, it was felt reasonable to conceptualize teaching activities on the basis of the general principles of teaching mathematics. As such, an attempt has been made to

¹ & ². Passi, B.K., loc.cit., p. 11-12
³. Pinsent, A., loc.cit., pp. 50-54 (chap. II)
specify general mathematics teaching principles on the basis of objectives, psychological theories and expert's opinion in this direction. In this connection a short discussion about the objectives and psychological theories for effective mathematics learning was felt essential.

4.2. **Objectives and Theories of Teaching Mathematics**

Since the present study has been delimited to classes III to VI, so in consideration of grade and developmental levels of school children, some of the important objectives of teaching mathematics are stated below:

1. To experience satisfactions that come to individuals possessing knowledge of related contents.

2. To develop abilities and interest in observing, defining, comparing and analysing.

3. To develop computation-skills and ability to estimate and verify answers.

4. To develop understanding of Mathematical vocabulary and structure of operations.

5. To develop positive attitude towards the subject and to create "motives" for continued study in higher level.

6. To develop "personality trait" useful in responsible citizenship.

All these objectives cannot be achieved in a single way but through several avenues by which learning becomes an
active process. Learning becomes active through reading, writing, listening, discussing, working, experimenting, drawing, observing, comparing, analysing, synthesising, interpreting and the like. But as Mathematics is a cumulative and a continuously expanding subject, it demands increasing thoughts in orchestrating these activities. Because, "unless a clearly and correctly formed Mathematical concept emerges from a learning situation", "the learning of all the later mathematics dependent upon this concept is seriously, sometimes totally, impaired." As for example, the lack of correct concept in fraction-equivalence may be one of the chief reasons for experiencing difficulty in fraction-addition or subtraction by majority of our school children.

To attain a maximum of usefulness of these learning activities, it is to be so planned that they can bring about desired changes towards understanding and mastery of the subject at pupils' levels of achievement. As generally conceived, these changes can be brought about through four fundamental instructional practices followed by some specific activities for each of them. The instructional practices are as follows:

(1) **Pre-Instructional Practice**: It involves testing of previous knowledge and pre-view of the new lesson, producing "behaviour changes", like recalling, comparing and motives for further learning.

---

(2) Instructional Practice: It involves creating challenging attitude, explaining through exploratory questions, guiding and directing, illustrating and use of examples, pupils' participation pupil-pupil interaction, pupil-teacher interaction and blackboard work producing "behaviour changes" like understanding and recognising.

(3) Closure (Teaching for assimilation and drill service): It involves testing, removing difficulties, reteaching, directing and guiding producing "behaviour changes" like recognising, understanding, interest (motives) and love for the subject.


(ii) Managerial: It involves teacher's order and suggestion, considerateness and classroom discipline.

4.3. Mathematics Teaching Principles (MTP)

For the present investigation, a few important general principles of teaching mathematics that are conceived to be desirable for effective mathematics learning have been collected from various books and journals. But as our study has been

---

delimited to classes III to VI, it was felt reasonable to select the principles in consideration of grade and developmental levels of our school children. As such, altogether thirty six principles for different phases of teaching have been tentatively selected. A detailed list is shown in the Appendix-C(i).

4.4. Conceptualization of Teaching Activities

In order to formulate and identify Teaching Activities, it was thought reasonable to explore and identify the actual classroom situations relevant to materials of the list stated above; and to isolate those items which could not be found functioning in the existing pattern of instructional programme. For this purpose, it was the plan to discuss with Mathematics Teachers (III-VI) of different schools in Kohima. At this initial stage, the proximity to Kohima was desirable because collection of data would be more facilitating and selection of schools would be diversified according to size and sex-sector variabilities.

Accordingly, ten schools* - three from town and seven from neighbouring villages were chosen. Altogether, 21 teachers of diversified qualifications and experiences were obtained from these ten schools as may be found from the Table 4 given below:

---


*Distributions of schools at Kohima are shown in Appendix-B(i)
Table 4: Showing Qualifications and Experiences of 21 Teachers selected from Kohima, 1975-76

<table>
<thead>
<tr>
<th>Area</th>
<th>No. of schools visited</th>
<th>G</th>
<th>I</th>
<th>M</th>
<th>UM</th>
<th>10-aa</th>
<th>9-5</th>
<th>below 5</th>
<th>Total No. of teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Urban</td>
<td>3(16)</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Rural</td>
<td>7(31)</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Total:</td>
<td>10(47)</td>
<td>2</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>21</td>
</tr>
</tbody>
</table>

Note: Years of experience were noted based on teachers' statement on the spot.

G = Graduate, I = Intermediate, M = Matriculate,
UM = Under Matric, aa = and above

By a careful discussion with the teachers it was revealed that mathematics teaching in actual classroom situation had been done on the basis of almost all the principles in various degrees except in cases of MTP having serial numbers* 4, 16, 18, 22, 25. These were, therefore, dropped out. Teaching activities were conceptualized on the basis of remaining 31 teaching principles. Altogether, 21 classroom teaching activities were conceptualized of which 17 were based mostly on "skills" developed at Sandford University and CASE (Baroda); and the remaining 4 activities were based on psychological backgrounds and personal experience in teaching mathematics. A detailed list of conceptualized teaching activities is shown in Appendix-C(ii).

*For detailed description, see Appendix-C(i).

MTP – Mathematics Teaching Principle
Rationalae for Conceptualisation of Teaching Activities based on personal experience:

It has been recognised that not all learning begins in concrete materials, but that many may start in "already learned abstractions". As for example, unitary method may be referred back to "abstractions" already learned in ratio proportion. This multiple approach-based teaching not only provides real meaning of pupils' previous experience, but also gives increasing interest for the subject - (CTA - 7).

While developing a new lesson, it is often necessary to anticipate probable difficulties. The points of difficulty need repeated emphasis, failure of which may result in blocking the understanding of subsequent parts of the unit - (CTA 9).

In consideration of grade (III-VI) and developmental levels, our students are supposed to be immature in co-ordinating their previous experiences necessary for the solution. Here teacher is - to build up and co-ordinate necessary background through "directed questions", to supply direct information where student is adrift, to encourage self checking for developing independent study habit - (CTA 19).

It is most frequently observed that students allow statements (explanation) to go unchallenged or uninterrupted even though they do not understand. Such barriers to freedom of inquiry should be broken down by sympathetic and kindly encouragement (CTA 20).
4.5. **Example of Questioning**

In consideration of the degree of importance of teaching activities in the field of methodology of teaching mathematics, "questioning technique" has been considered as the most powerful weapon. In the pre-instructional practice, it can ascertain whether pupils' background is adequate to make the new lesson intelligible and acceptable. In instructional practice, "judicious questioning" in an exploratory manner can bring to light weak points of pupils' knowledge as well as helps to increase pupils' participation. Santhanam has pointed out that the "quantum of teacher questioning in mathematics has been more than in any other school subject."

It was, therefore, considered useful to give an example to illustrate "questioning" technique in mathematics teaching.

**Example:** It is assumed that pupils know how to find the area of a rectangular plane surface by calculating total number of square units in it (graphic method). But they do not know the mathematical formula for it (i.e., Area = length x breadth). Pupils are asked to draw a rectangle having, say, 6 inches length and 4 inches breadth; and to find the area by counting the total number of square inches in it.

---


Questions:

Teacher (T): You are going to examine this process in deriving a shorter process for determining the area of a rectangle. Can any of you think of a shorter way?

Pupil (P): Yes, by counting sq. inches in one row and multiplying it by number of rows.

T: Good, but it is not short enough, because you are to draw the rectangle and divide it into square inches. Think for another way.

P: No response.

T: Without actually counting, can anybody of you tell me how many such squares would be in one row?

P: No response.

T: What have you done in dividing the rectangle into square inches?

P: Length and breadth are divided into inches.

T: Good, now can you tell me the number of divisions you will get along the length?

P: 6 divisions.

T: What would be the division if length is 25 inches?

P: 25 divisions.

T: Then, what will be the number of divisions for a certain length?

P: As many divisions as there are inches in it.
T : What will be the number of rows? (without actually counting)

P : 4 rows

T : What will be the number of rows for a certain breadth?

P : There are as many rows as there are inches in breadth.

T : Very good: Now tell me what will be the total number of such squares in a rectangle having certain length and breadth?

P : Number of inches in length into (x) number of inches in breadth.

4.6. Experts' Opinion

Though it was mentioned in the preceding section that majority of the CTA's had been conceptualised based mostly on skills developed at CASE and Standford University, yet many of them had to be modified in consideration of the nature of the subject* as well as grade and developmental levels of the students involved in our study. Moreover, 4 CTA's were conceptualised on the basis of personal experience in the line. It was, therefore, reasonable to justify the consistency and "content validity" in respect of these activities.

Fourteen judges were selected from the Institutes of Education, general College and Higher Secondary Schools in Nagaland, Manipur and Tripura. The proximity to Nagaland was obviously desirable not only that personal approach would be more economic but also to maintain homogeneity in many respects such as backwardness, food habits, social structure, educational

* ref. chap. 2
background and the like. The selection of judges was based on academic qualification, professional qualification, experience, interest in the field and willingness to co-operate in such a study of our short. A list of judges with their qualifications and experiences is shown in Appendix-C(iii).

A copy of the list of selected CTA's as shown in Appendix-C(ii) was sent to each of the fourteen judges along with specific instructions regarding how to judge and what to judge. A copy of the detailed instructions is enclosed in Appendix-C(iv).

It took a long time, nearly three months to receive judges' ratings. Four judges who had originally agreed to co-operate, could not complete even within this time. However, out of fourteen judges, the ratings from ten judges were received within this time. As the instructions were clear and all judges were qualified and experienced in this line, it was assumed that a dependable conclusion could be drawn through the ratings of these (10) carefully selected judges. Appendix-C(v) shows the summary of ratings by ten judges.

4.7. Consistency of Judges' Ratings

In order to determine the consistency of response of a judge over components (items) the investigator could not find a method better than the method of "one-way repeated measure of ANOVA". Since we had repeated measures on the same group
of judges, the "interaction" of judges and components would be a suitable measure of inconsistency.¹

If the judgment are consistent and effectiveness of components are well discriminated, then the variance due to judges (more accurately "interaction") must be very small in comparison with the variance due to components. The details of analysis are shown in the Table 5. It is observed that the variance due to judge x component is very small (0.578) compared to the variance due to components (3.375). This indicates real difference ($F = 5.84$, df $20/180$, $P < .01$) between components. On the other side the variance due to judges is not found to be significant at $P = .05$. It is, therefore, concluded that there exists a consensus among different judges regarding efficiency of the CTA's.

4.8. The Observation Schedule for MTCS

Being encouraged with consensus of experts, an observation check list consisting of 21 components (activities) under four major heads, viz., (1) creating set, (2) development, (3) closure, (4) managerial has been developed. A specimen copy is enclosed in the Appendix-C(vi).

This check list is meant to ascertain the degree to which the teacher to be rated exhibits each of the 21 activities

Table 5: ANOVA of Judges' Ratings  
(Data in Appendix C(v))

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Sums of Squares (SS)</th>
<th>df</th>
<th>Estimate of variance (Mean square)</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among the means of components</td>
<td>67.4953</td>
<td>20</td>
<td>3.375</td>
<td>5.84**</td>
</tr>
<tr>
<td>Within groups</td>
<td>115</td>
<td>189</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between means of Judges</td>
<td>10.9714</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Component X Judges</td>
<td>104.0286</td>
<td>180</td>
<td></td>
<td>.578</td>
</tr>
<tr>
<td>Totals</td>
<td>182.4953</td>
<td>209</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** = Significant at \( P \leq .01 \)
in actual classroom situation. Judgement has to be made on a seven point scale ranging from "0" to "6". The scale value "0" against an activity indicates that the concerned activity was not at all exhibited by the teacher in the class, while the scale value "6" indicates that it was exhibited very extensively by the teacher. Keeping these two extreme points of the scale in view the judgement was given by encircling the number in the scale provided in the right hand side of the check list.

4.9. The Methods of Observation

Various procedures for observing classroom teaching behaviours have been discussed by Dr B.K. Passi. They are given below:

(i) by coding category numbers after the occurrence of corresponding pupil-teacher behaviour at the end of a fixed time interval.

(ii) by entering tallies against the category after their occurrence at the end of fixed time interval, and

(iii) by giving rating at the end of lesson against each category of behaviour.

In our case, though the last approach was employed, yet the tally system was also used only by the investigator for "interaction analysis".*

1. Passi, B.K., loc.cit., p. 272

*for details, ref. chap. 5
Each classroom observation was made independently by the investigator and another qualified observer. This was done in order to determine observer's reliability which would be presenting in the following pages. A stratified sample of 20 schools* was visited for this purpose. Altogether 49 classrooms, one for each teacher, were observed by the investigator out of which paired observations were made simultaneously on 31 different classrooms. In each school, the marked observation check-list were collected from the second observer on the same day.

4.10. Observer-Reliability

The objectivity of a scale is another way of testing reliability of the scale. Reliability in terms of objectivity was studied, by obtaining the degree of agreement (correlation method) of two observers (the investigator being one observer) who made independent ratings for each teacher in the same classroom. It was decided to select one seniormost qualified mathematics teacher** preferably from the higher class section of the schools. 15 such schools were found suitable for this purpose. The interested teachers of attached senior section were requested to act as judges (observers) in the study. This approach was made personally on the day of administering the

*List of schools visited is presented in Appendix-B(iii)

**Academic qualification ranges from I.Sc. to B.Sc. (with Mathematics)
PAI Test. The teachers who were found really interested in this field and voluntarily agreed to act as judges were selected.

Each observer was given necessary and sufficient instructions regarding how to judge and what to judge. Each of them was also supplied with a copy of observation check-list together with glossary of key-words.* Ratings by senior teachers are shown in Appendix-C(viii).

Reliability by Correlation Method

The group of observers (here senior teachers) could reasonably be considered homogeneous, as all were given the same instructions, and also they came from the same professional group having more or less same academic qualification. Based on this assumption, the reliability of rating scores could be estimated by correlating one set of scores (rated by the investigator) with another set of scores (rated by other observer).

Altogether, a sample of 49 classrooms were observed. Out of these, 31 paired observations were made independently. Paired scores for each classroom were correlated. The co-efficient of correlation was computed by using the formula \( r = \frac{N \Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{N \Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{N \Sigma y^2 - (\Sigma y)^2}} \)

\*Ref. Appendix-C(vii)

The reliability co-efficient is found to be $r_{ii} = 0.77$ for $N = 31$. Table 6 gives Pairs of Observation Ratings for 31 teachers.

The reliability co-efficient obtained is obviously highly significant, thus indicating an well acceptable degree of agreement between investigator's ratings and other observer's ratings. This result well agrees with the result found by Wrightstone\(^1\) in his study, "Measuring the Social Climate of a Classroom". He obtained $r = 0.92$ for a sample of 25 teachers. On the basis of above evidence, our instrument could be used with high degree of dependability to measure mathematics teaching success through classroom observation.

\* The Spearman-Brown formula was applied for determining reliability of ratings by two persons.

\(\Delta\) Garrett, H.E., loc. cit. p 343

Table 6: Showing Paired Scores on Observation Ratings for 31 Teachers (Data in Appendix C(viii), G)

<table>
<thead>
<tr>
<th>Sl.No.</th>
<th>Case No.</th>
<th>Investigator (X)</th>
<th>Other observer (Y)</th>
<th>Sl.No.</th>
<th>Case No.</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>93</td>
<td>78</td>
<td>16</td>
<td>20</td>
<td>52</td>
<td>65</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>48</td>
<td>63</td>
<td>17</td>
<td>21</td>
<td>38</td>
<td>53</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>74</td>
<td>91</td>
<td>18</td>
<td>22</td>
<td>79</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>47</td>
<td>60</td>
<td>19</td>
<td>24</td>
<td>34</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>43</td>
<td>64</td>
<td>20</td>
<td>25</td>
<td>54</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>35</td>
<td>52</td>
<td>21</td>
<td>27</td>
<td>64</td>
<td>47</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>81</td>
<td>77</td>
<td>22</td>
<td>29</td>
<td>54</td>
<td>69</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>33</td>
<td>49</td>
<td>23</td>
<td>35</td>
<td>54</td>
<td>55</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>80</td>
<td>79</td>
<td>24</td>
<td>36</td>
<td>71</td>
<td>57</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>69</td>
<td>52</td>
<td>25</td>
<td>37</td>
<td>55</td>
<td>63</td>
</tr>
<tr>
<td>11</td>
<td>15</td>
<td>62</td>
<td>47</td>
<td>26</td>
<td>39</td>
<td>62</td>
<td>65</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>39</td>
<td>43</td>
<td>27</td>
<td>40</td>
<td>39</td>
<td>65</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>67</td>
<td>54</td>
<td>28</td>
<td>41</td>
<td>49</td>
<td>73</td>
</tr>
<tr>
<td>14</td>
<td>18</td>
<td>39</td>
<td>37</td>
<td>29</td>
<td>42</td>
<td>42</td>
<td>54</td>
</tr>
<tr>
<td>15</td>
<td>19</td>
<td>47</td>
<td>39</td>
<td>30</td>
<td>43</td>
<td>40</td>
<td>41</td>
</tr>
</tbody>
</table>

Calculation:

\[
\Sigma x = 1691, \quad \Sigma y = 1779, \quad \Sigma xy = 101116 \\
\Sigma x^2 = 99901 \quad \Sigma y^2 = 107705, \\
\]

\[
r_1 = 0.621, \quad r_{11} = 0.77
\]
Headmaster's Rating-Scale for Teacher Behaviour (HRTB)

4.11. Selection of Items:

"Behaviour traits" are generally used to evaluate student-teachers. J.A. Bond has presented 32 "traits" which had been in use for several years in U.S.A. to evaluate student-teachers at all levels. Among them 15 important "traits" have been selected for the purpose of using them as "pseudo criteria" in measuring teaching. The "traits" are as follows:

1. Co-operation
2. Job satisfaction
3. Job interest
4. Job responsibility
5. Emotional adjustment
6. Initiative
7. Mathematical attitude
8. Adaptability of teaching
9. Instructional judiciousness
10. Innovativeness
11. Adaptability to novelty
12. Disciplinary ability
13. Classroom sociability
14. Self improving attitude
15. Organising ability (of "Teaching Skills").

4.12. Experts' Opinion

Three experts* were chosen from Kohima because they were easily available. Among them, two were selected from


*Ref. Appendix-C(iii)
Nagaland College of Education, and one from Science College, Kohima. A detailed discussion was made with each of them independently to know how effectively the selected traits could be used as "Pseudo Criteria" of measuring classroom teaching in Mathematics. All the three experts ensured the investigator that the inventory, as a whole, could be considered adequately valid. Being encouraged with satisfactory "content validity", the investigator proceeded to further step.

4.13. Data Card

A bio-data card was included in the inventory to obtain some more details of mathematics teachers (III-VI). Qualification and teaching experiences were specially included to ascertain the degree of relationship between teaching success and teacher characteristics (detailed discussion was made in Chapter 3). Necessary instructions were also given. A specimen copy of this inventory is given in Appendix- C (ix).

4.14. Scaling

Compared to the Thurstone-type and the Rammer-type scales, the Likert-type scales are fairly easy to construct and so it was employed in constructing our HRTB with the assumption that the behaviour "traits" are normally distributed. Accordingly, for each item, a five point scale: Very Low, Low, Average, High, and Very High was introduced. The scales were scored by assigning the values from 1 to 5, 1 being the unfavourable end and 5, the
favourable end of the response continuum. The total values were considered as a teacher's score.

4.15. **Administration of the Questionnaire**

The administration of this inventory depended entirely on the voluntary consent of the headmasters of selected schools. The specimen copies of the schedule were distributed to headmasters on the spot with a few minutes talk explaining the purpose in brief. All the headmasters, except one, readily agreed. They were requested to respond to individual items on the strength of their adequate information from various sources and day-to-day observation. Any confusion raised for any item was cleared up on the spot. The filled-in questionnaires were collected back on the same day. Altogether, ratings for 45 teachers (out of 49) were obtained.

4.16. **Reliability of Ratings**

Two methods, viz., "split-half" method and "Test-Retest" Method were employed. Table 7 presents the pairs of scores on odd items (X) and even items (Y). The self correlation of the half test was estimated by PPMF$^1$ and $r_1 = 0.76$ was obtained. The self correlation of the whole test was then estimated by SBPP$^2$ and $r_{44} = 0.86$ was obtained. It is found to

2. ibid., p. 339 (formula-79)
Table 7: Showing Paired Scores on Odd and Even Items of HRTB for 45 Teachers
(Data in Appendix H)

<table>
<thead>
<tr>
<th>Cases</th>
<th>X</th>
<th>Y</th>
<th>Cases</th>
<th>X</th>
<th>Y</th>
<th>Cases</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26</td>
<td>24</td>
<td>16</td>
<td>23</td>
<td>22</td>
<td>31</td>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>18</td>
<td>17</td>
<td>24</td>
<td>20</td>
<td>32</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>29</td>
<td>18</td>
<td>19</td>
<td>19</td>
<td>33</td>
<td>22</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>22</td>
<td>19</td>
<td>18</td>
<td>19</td>
<td>34</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>17</td>
<td>20</td>
<td>19</td>
<td>19</td>
<td>35</td>
<td>29</td>
<td>24</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>18</td>
<td>21</td>
<td>26</td>
<td>25</td>
<td>36</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>14</td>
<td>22</td>
<td>25</td>
<td>18</td>
<td>37</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>20</td>
<td>13</td>
<td>38</td>
<td>24</td>
<td>22</td>
</tr>
<tr>
<td>9</td>
<td>31</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>18</td>
<td>39</td>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>12</td>
<td>25</td>
<td>28</td>
<td>24</td>
<td>40</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>17</td>
<td>26</td>
<td>24</td>
<td>21</td>
<td>41</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>12</td>
<td>23</td>
<td>21</td>
<td>27</td>
<td>31</td>
<td>25</td>
<td>42</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>22</td>
<td>21</td>
<td>28</td>
<td>19</td>
<td>17</td>
<td>43</td>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>14</td>
<td>19</td>
<td>22</td>
<td>29</td>
<td>23</td>
<td>20</td>
<td>44</td>
<td>23</td>
<td>16</td>
</tr>
<tr>
<td>15</td>
<td>17</td>
<td>14</td>
<td>30</td>
<td>20</td>
<td>15</td>
<td>45</td>
<td>19</td>
<td>20</td>
</tr>
</tbody>
</table>

\(X = \text{aggregate score on odd-item, } Y = \text{aggregate score on even item.}\)

Case No. corresponds to serial No. in Appendix H.

**Calculation:**

\[
\begin{align*}
\Sigma x &= 1029, \quad \Sigma y = 892, \quad \Sigma xy = 20853 \\
\Sigma x^2 &= 24197, \quad \Sigma y^2 = 18220, \\
\frac{r_1}{\hat{r}} &= 0.76, \quad r_{11} = .86
\end{align*}
\]
be highly significant ($P < .01$).

For "test-retest" reliability, the same rating schedules were mailed to headmasters of all the selected schools after about a month of first occurrence. Ratings were obtained from 16 schools involving 36 teachers. The time break between the first rating and second rating ranged from one and half months to two months. Table 8 presents the paired scores on first performance and second performance. The co-efficient of correlation was computed by PPMF and $r = 0.82$ was obtained.

The high degree of "odd-even" reliability co-efficient could be explained by the tendency of headmasters to give similar ratings on 'traits' that seemed to them logically related. Cook and Carroll in measuring teachers' personality obtained $r = 0.87$ as an "odd-even" reliability. In support of this evidence, our result could be well accepted.

Table 8: Showing Paired Scores on HRTB at Two Different Intervals

<table>
<thead>
<tr>
<th>Sl. No.</th>
<th>X</th>
<th>Y</th>
<th>SI. No.</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>59</td>
<td>19</td>
<td>33</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>38</td>
<td>20</td>
<td>43</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>62</td>
<td>67</td>
<td>21</td>
<td>52</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>48</td>
<td>42</td>
<td>22</td>
<td>45</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>38</td>
<td>41</td>
<td>23</td>
<td>56</td>
<td>40</td>
</tr>
<tr>
<td>6</td>
<td>41</td>
<td>46</td>
<td>24</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
<td>44</td>
<td>25</td>
<td>43</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>43</td>
<td>40</td>
<td>26</td>
<td>35</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>54</td>
<td>58</td>
<td>27</td>
<td>43</td>
<td>47</td>
</tr>
<tr>
<td>10</td>
<td>28</td>
<td>32</td>
<td>28</td>
<td>53</td>
<td>53</td>
</tr>
<tr>
<td>11</td>
<td>37</td>
<td>42</td>
<td>29</td>
<td>42</td>
<td>36</td>
</tr>
<tr>
<td>12</td>
<td>44</td>
<td>55</td>
<td>30</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>13</td>
<td>43</td>
<td>43</td>
<td>31</td>
<td>53</td>
<td>57</td>
</tr>
<tr>
<td>14</td>
<td>41</td>
<td>49</td>
<td>32</td>
<td>40</td>
<td>37</td>
</tr>
<tr>
<td>15</td>
<td>31</td>
<td>38</td>
<td>33</td>
<td>48</td>
<td>51</td>
</tr>
<tr>
<td>16</td>
<td>45</td>
<td>44</td>
<td>34</td>
<td>48</td>
<td>43</td>
</tr>
<tr>
<td>17</td>
<td>51</td>
<td>61</td>
<td>35</td>
<td>36</td>
<td>32</td>
</tr>
<tr>
<td>18</td>
<td>43</td>
<td>49</td>
<td>36</td>
<td>41</td>
<td>44</td>
</tr>
</tbody>
</table>

N.B.  \( X = \) Scores of first ratings,
      \( Y = \) Scores of second ratings.

Calculation:

\[
N = 36
\]

\[
\sum x = 1561, \quad \sum y = 1641, \quad \sum xy = 92924
\]

\[
\sum x^2 = 69619, \quad \sum y^2 = 77211
\]

\[
r = 0.819
\]
Pupil Attitude Inventory (PAI) for Evaluating Mathematics Classroom Teaching

In the preceding section of this chapter, the teaching activities that have been conceived for effective mathematics learning were thoroughly examined and an observation schedule consisting of 21 teaching activities was developed. The present section deals with details of construction and standardisation of rating scale of pupils' attitude (opinion) towards how such teaching activities have been exhibited or practised by the teacher in actual classroom situations.

4.17. Methods of Rating Scale Construction

The various techniques for constructing rating scale have been discussed by Freeman. They are as follows:

1. **Critical Incidental Technique**: This involves detailed description of individual behaviour to be regarded as favourable or unfavourable in a given situation.

2. **The Q-Sort Technique**: Here the rater is usually given a large number of "traits" to sort out according to the degree to which he regards each "trait" as desirable or undesirable in persons with whom he must associate.

---

(3) The Forced Choice Technique: Here the respondent is asked to indicate which one (or two) of the statements in each set consisting of two or more items, is most appropriate to the persons being rated. The widely used form under this method is the traditional 'Yes' and 'No' form of checklist. But in this type of items "many subjects experience difficulty in choosing a response alternative" because their answers appear to fall between the clearcut 'Yes' and 'No' response.¹ In view of this disadvantage, a more discriminating type of checklist is devised which provides the respondents with an opportunity to indicate the degree to which they believe each trait is manifested by the subject being rated. This is usually done by adding to each statement a series of qualifying terms (words), such as always, often, frequently, usually, sometimes, rarely, seldom, never and the like.

The first method requires competence and insights and the second one conscientious efforts on the part of the raters. But the present inventory deals with pupils of grades (classes) III to VI and so, could not be treated as sufficiently mature to express their opinions more critically. The first two methods were, therefore, rejected and the third one was employed. The advantages of the 3rd technique are as follows:

(i) It reduces possible effects of biases in either direction.

2. Freeman, F.S., loc.cit., pp. 527-528
(ii) It yields more valid results than any other method.

(iii) Statements are kept within a narrower range of standards thus providing a unified concept of rating and "halo effects" do not operate so strongly as with other methods.

To ensure more valid results, the work was planned as a three phase activity, namely, tryout phase I (pre-tryout), tryout phase II (Inter-tryout) and experimental phase III (Final form). But in order to minimise labour, "Yes - No" form of checklist was used in the preliminary stage. This was followed by "forced choice type" of multiple alternatives in the final form. The exact number of choices was studied empirically before selecting items for final form.

4.18. Methods for Collecting Opinions

The usual methods for collecting opinions as suggested by Thurstone and Chave¹ are -

(i) to ask a large group of individuals to write out their opinions about the topic under consideration.

(ii) to consult the current literature for suitable brief statements.

In addition to these two methods, another method is -

(iii) to collect opinions from personal interview and discussion with persons whose attitudes are intended to be measured.

But as the present inventory deals with pupils of classes III-VI, it could be reasonably assumed that they would not be sufficiently mature to express their feelings critically about classroom behaviours of their teachers. Moreover, it might involve subjective preferences to a great extent. The first method was, therefore, left out; and the second and third methods were employed. The interview method (3rd method) was employed to have natural form of opinions which, in turn, would enable elimination of artificial responses to a great extent and thus increasing validity of opinions.

4.19. Tryout Phase I (Pre-Tryout)

Based mainly on conceptualised teaching activities as described in earlier section of this chapter, an interview schedule was prepared (see Appendix-D(1)). To facilitate interview work, all the 21 teaching activities were classified under 7 different aspects and altogether 13 questions were designed. They are shown below:

<table>
<thead>
<tr>
<th>Aspects</th>
<th>No. of questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Creating motives</td>
<td>2</td>
</tr>
<tr>
<td>2. Asking questions</td>
<td>2</td>
</tr>
<tr>
<td>3. Explaining</td>
<td>3</td>
</tr>
<tr>
<td>4. Pupils participation and encouragement</td>
<td>2</td>
</tr>
<tr>
<td>5. Guiding and directing</td>
<td>2</td>
</tr>
<tr>
<td>6. Considerations and emotional stability</td>
<td>1</td>
</tr>
<tr>
<td>7. Creative ability</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>13</strong></td>
</tr>
</tbody>
</table>
Ten teachers were chosen from five different schools at Kohima town and the interview was taken with five students at random for each teacher. It was assumed that students at these levels (III-VI) would be sufficiently mature to respond to selected questions to various degrees of favourableness or unfavourableness. The co-operation of a qualified local Sub-Inspector of schools was taken to help in communicating the key concept of the situation, where necessary through local dialects. The purpose of the interview was explained to the students and they were given assurance that the response would be kept most confidential. In order to increase pupils' confidence to a greater extent, the concerned Headmaster was requested to give complete assurance to his pupils in this respect. The Headmaster was assured that the identity of the teacher would not be divulged. The precautions undertaken in this manner were considered enough for expressing their feelings with complete freedom.

4.20. **Classification of Responses**

The responses of the interview were scrutinised for examining the scalability and ultimately classified into four different groups (categories) in order of their degrees of favourableness. The classification was made in such a manner that each response in a category expressed identical degree of favourableness.

The opinions of the experts, in this classification, could not be considered essential because it was based on empirical method. It was hoped that the judgement undertaken
by the investigator himself would provide sufficient indication of the spread of responses in attitude continuum. The four groups were represented by the letters a, b, c, and d in order from extreme favourable to extreme unfavourable ends in case of positive statements and reversed in case of negative statements. The classified groups consisted of only catch words.*

4.21. Scalability of Responses

Before proceeding further, it was considered useful to justify the scalability of the responses made by the pupils in the interview. This was done by drawing bar graphs of proportions at each response category of various teaching aspects. Table 9 presents the proportions.

Table 9: Showing Proportions of Interview Responses at various categories made by 50 students for 10 teachers (5 each)

<table>
<thead>
<tr>
<th>Teaching aspect</th>
<th>Response in categories</th>
<th>Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>1. Creating motives</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>2. Asking questions</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>3. Explaining</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>4. Pupils' participation and encouragement</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>5. Guiding and directing</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>6. Considerateness and Emotional stability</td>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>7. Creative ability</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

* for details ref. Appx - D (ii)
Corresponding graphs (Ref. Chart 2) were drawn to examine the spread of responses. Graphs for items 1 and 5 indicate, to some extent, decreasing tendency towards unfavourable ends. Only the graphs for items 4, 6, and 7 indicate some form of normal distribution. In few cases, bimodal distributions were observed. Taken together, all graphs justified that the responses to various teaching aspects were very complex and, therefore, Likert type or Remmer type scaling could not be made use of; and so, it was planned to employ the method of "empirical keying" for scaling of responses of the final questionnaire.

Applying the criteria (ref. Appendix-D(iii)) for formulating statements as suggested by Thurstone, Wang, and Ojemann, a questionnaire consisting of thirty seven statements of "Yes - No" form, covering all the twenty one activities, was prepared for the Tryout Phase II. A few statements having more or less the same concept were deliberately introduced in different ways in order to isolate the best fitting item to be used in the Final Form. Selection of too many statements was also avoided from psychological consideration. These procedures would ensure a high degree of validity of items in respect of sample under consideration. A specimen copy of Inter-Tryout questionnaire is enclosed in Appendix-D(iv).

CHART 2  Bars Showing Spread of Interview Response (N=50)  
(Data of Table 9)
The objective criterion of the present inventory is to determine how well the statements can discriminate teachers in respect of classroom teaching behaviours. For this purpose, two criterion groups of teachers - superior group and inferior group were selected on the basis of Headmasters' opinion. Ten schools (out of 47) located at various places of Kohima were visited. Table 10 presents distribution of schools in Kohima and number of criterion teachers selected.

Table 10: Showing Distribution of Schools (Kohima, 1975), number of schools visited and teachers enlisted in criterion groups

<table>
<thead>
<tr>
<th></th>
<th>Urban</th>
<th>Rural</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Govt.</td>
<td>Private</td>
<td>Total</td>
</tr>
<tr>
<td>No. of schools</td>
<td>7</td>
<td>9</td>
<td>16</td>
</tr>
<tr>
<td>Schools visited</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Teacher available</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Teachers in Criterion Groups</td>
<td>1*</td>
<td>1*</td>
<td>2*</td>
</tr>
</tbody>
</table>

Note: * indicate superior teacher, **indicate inferior teacher.

Source: Directorate of Education, Government of Nagaland, Inspection Branch, 1975

Altogether 7 teachers (out of twenty one) were included in the criterion groups, four in superior and three in inferior group. In order to facilitate test administration, it was decided to select 10-12 students at random from a single class for each teacher.
4.23. Procedure for Administering Test (Inter Tryout)

Since students had to express their feeling about their teachers, their confidence had to be won first. This was done by (i) impressing upon the student's needs and purpose of the investigation and (ii) giving assurance for secrecy of their responses by the proper authority. Being ensured of their frank responses, the questionnaires were distributed to the students of selected classes. They were given proper instruction for giving their reactions to each item without leaving a single statement blank. No time limit was given. But they were asked to react as quickly as possible. While conducting the test, nobody was allowed to enter into the room except the investigator himself. The marked questionnaires were taken back after the completion of the test. All these activities ensured better security on the part of students.

4.24. Selection of Statements for Final Form

The usual method known as "Criterion Method" was employed at this stage. Accordingly, the answered questionnaires in respect of criterion teachers (4 in superior and 3 in inferior) were sorted out in two piles, one for each criterion group. Altogether, 84 students under 7 criterion teachers (12 for each) were obtained among which a few papers in both the piles were found incomplete and were rejected. However, about 80% (67 out of 84) of the obtained questionnaires were found acceptable. The accepted data are shown in Appendix-D(v). They were subjected to
analysis for final selection on the basis of the following considerations:

To ensure greater dependability of a rating scale, it is essential that the variability within pupils should be reasonably small. But in our case, the individual differences among teachers, in each criterion group could not be avoided and so, a high degree of agreement among the raters in each group could not be expected. Moreover, we could not expect a sharp discrimination between the two criterion groups in respect of all the significant items, although items having high degree of discriminative power indices are usually accepted for such type of tests. However, the degree of acceptance is, to some extent, arbitrarily determined and there is no rigid principles as to what degree of discriminative power index should be necessary.¹

As our main aim is to measure teaching behaviours, so an item which can well differentiate the functions of CTA's between these two criterion groups of teachers is to be chosen. An item which fails to differentiate is virtually regarded as constant and so relative measure cannot be made with such type of item. Items of this nature are, therefore, justified to be discarded.

In view of above considerations, it was planned to select items on the basis of the following criteria:

(i) Average acceptance - lying between 20% and 80%.

¹ Freeman, loc.cit., pp. 112-117

CTA = Class-room Teaching Activity
(ii) Significant discriminative power ($P = .05$)

(iii) Relevancy of items (Graphic method).

Accordingly, the proportions of favourable acceptance (here defined as "Popularity Index" = PI) for each criterion group were calculated and the average PI's were estimated for each item by the usual method. In measuring "discriminative power", it is the usual procedure to employ $\chi^2$, $\chi^2$-test, especially, when criterion groups are used and measurements are dichotomized. Kilpatrick$^1$ used $\phi_c$, Devault & Cheong,$^2$ Cook and Carroll$^3$ and many others used $\chi^2$ as recommended by McNemar. For the present test, since measurements were not continuous rather dichotomised, $\phi_c$ was considered more suitable and simple. $\chi^2$ was first calculated and then converted into corresponding $\phi_c$ by the formula:$^4$

$$\chi^2 = N\phi_c^2$$

Table of Appendix-D(vi) shows the summary of results for $^0I$, $\phi_c$, and level of significance (at $P = .05$ and .01) in respect of 37 items.

---


The table reveals that twenty items are significant even beyond .01 level and five items at .05 level of which three items, viz., item Nos. 22, 27 and 33 do not satisfy the criterion (i) and so they are to be discarded or modified. But before doing so, criterion (iii) is to be employed. Thurstone used "index of similarity" based on method of similar reactions. According to him, when two or more "traits" tend to co-exist, they are regarded as functionally similar; and when they are more or less mutually exclusive, they are regarded as functionally dissimilar. The degree of similarity can be measured by $\phi_c$ co-efficients. But this is a rigorous process logically sound. It would involve $\frac{25 \times 24}{2}$ or 300 pairs of comparison for 25 significant statements. Also "point biserial correlation" ($r_{p.bis}$) could not be used because criterion groups were employed.

It was found convenient to use a common basis of comparison of PI's by graphic method. Accordingly, data of Appendix-D(vi) were utilised. The average of all the average PI's was computed and indicated in the graph by a thick line. Deviations of each individual PI were marked by vertical lines. The details are shown in Chart 3.

Deviations should be reasonably small when activities are functionally similar. Based on this rationale, it could be observed that items 9, 22, 27, 28 and 33 tended to be partially irrelevant. So, in consideration of criteria (i) and

CHART 3  Deviation Diagram of PAI Inter-Try-out Consistant Items (25 numbers) (Data of Appendix D(vi))
(iii), these items had to be either eliminated or modified in the final stage.

4.25. Experimental Phase III (Final Form)

From the inter tryout questionnaire, altogether we have selected 20 items for the construction of PAI-Final Form. But test of this sort needs revision for several times before it is used for research work. The considerations made for this purpose were the following:

(1) To re-examine brevity and clarity of expression,
(2) To re-examine if any duplication occurs,
(3) To re-examine poorly stated items, if any,
(4) To re-examine whether all the teaching activities were covered by the selected items.
(5) To re-examine if any important activity (or activities) was overlooked.

A statement distribution diagram (Chart 4) was drawn to show significant and non-significant items together with average popularity index (API) corresponding to individual CTA's. Items were marked on the horizontal line across each CTA. From a thorough scrutiny of the questionnaire (Inter tryout) and by a careful examination of this diagram the following things were observed:

(1) Statements expressing the same activity, that were deliberately introduced for isolating best fitting statement, were loaded with different degrees of $\Phi_c$. 
CHART 4  Distribution Diagram of PAI-Inter
Try-out Items (37 numbers) showing
Correspondence with CTA's, PI's & \( \phi \)’s
(Data of Appendix D(vi) )
(2) A few statements were over rated.

(3) Hardly 15 CTA's could be covered up by the items (20 numbers) selected so far.

(4) The activities (CTA's) that could not be covered up were the following:

(a) Use of teaching aids (items 28 ⇒ CTA-6)
(b) Emphasis on important steps (items 14 & 33 ⇒ CTA's 9 & 15)
(c) Black board work (item 27 ⇒ CTA-14)
(d) Home work (item 22 ⇒ CTA 17)
(e) Discipline (items 13, 30 and 34 ⇒ CTA 21).

In view of theoretical importance of these teaching activities, the partially irrelevant items (though statistically significant), viz., 22, 27, 28 & 33 had to be re-considered. The selection of items for a test of teaching evaluation is so complicated (since teaching process is a complex phenomenon) that theoretical considerations had to be given due weight besides other criteria. Besides, it appears that a few teaching activities are so interlinked that their individual identification cannot be made by students of grades III-VI. For this reason it was felt necessary to edit the statements. While editing, the following points were kept in view to ensure greater reliability and validity of the test as a whole:

(a) to modify statements expressing social desirability or undesirability, because there is a general tendency

⇒ stands for the sign of correspondence.
to over rate friends and teachers in respect of "traits" socially desirable.

(b) to modify statements expressing unfamiliarity of behaviours with the raters for facilitating conscientious ratings.

Based upon these considerations and with reference to the plan of the work (ref. Chap. 3), fifteen statements in the form of multiple choice (4 choices) were prepared. A specimen copy of English version is enclosed in Appendix-D (vii).

Table 11 showing correspondence between PAI test items and CTA's is presented below for checking at a glance.

To avoid personal bias, the items of the Final Form were randomly arranged. The English version of the Inventory thus prepared was translated by three qualified translators into three main Naga dialects, namely, "Angami", "Ao" and "Chang" corresponding to three selected zones in order to overcome the language difficulty. This ensured greater confidence about pupils' understanding of the key concept of item.

4.26. Instructions

Necessary instructions regarding how to mark the questionnaire were given in the first few pages of the booklet. Procedures were illustrated with suitable examples. The investigator was confident that such instructions were sufficient to enable pupils to indicate their feelings. In case of any confusion, students were asked to take help of the administrator (here the investigator himself).
Table 11: Showing correspondence between PAI test items and CTA's

<table>
<thead>
<tr>
<th>Final item</th>
<th>CTA's</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>13, 21</td>
</tr>
<tr>
<td>7</td>
<td>9, 10</td>
</tr>
<tr>
<td>8</td>
<td>15, 16</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>8, 14</td>
</tr>
<tr>
<td>13</td>
<td>17, 18</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>15</td>
<td>1, 2</td>
</tr>
</tbody>
</table>

Total: 15  21
4.27. Estimated Time

As the present inventory was intended for the pupils of intermediate grade group (grades III to VI), it was felt essential to give due consideration to the pupil’s attention span. The inventory was, therefore, kept restricted to 15 items. In many other studies of similar nature, it could be found that 15 to 20 items were considered enough for such grade group of students. It was estimated (based on tryout stage II) that hardly 20 minutes might be necessary to answer all the items. It was hoped that within this stipulated time, students could be able to give their conscientious ratings.

4.28. Sample of Schools Used in the Study

A stratified sample of schools was considered essential for better representativeness. The details were already discussed in the Chapter 3 and so, unnecessary repetition was avoided. The administration of this test depended entirely upon the voluntary consent of the Headmaster of concerned schools. Forcefulness was avoided to ensure greater reliability of the test. A copy of APPEAL expressing the needs and importance of such a study was mailed to the Headmaster of each of the selected schools. After a considerable period, twenty schools (about 67%) confirmed their willingness to co-operate in the present investigation. Since sampling was based absolutely upon stratification, the above

number could be considered adequate for the generalisation. A copy of proposed tentative programme was sent to each of the 20 schools and accordingly, they were visited.¹

4.29. **Selection of Qualified Student**

The reliability and the validity of ratings are usually dependent upon having a sufficient number of qualified judges.² Students of grades III to VI having adequate acquaintance (regularity being the indicator) with the teacher, were reasonably assumed to be qualified in this respect.

Though five to seven judges are frequently recommended for such purpose, yet the exact number of pupils (raters) could not be predetermined because number of regular students in a normal class could not be reliably pre-estimated. However, it was felt convenient to take minimum number of students to overcome administrative difficulties and some other undesirable activities, such as, peeping, discussing with fellow students etc. that were supposed to influence the test results significantly. Based upon these considerations, 12 to 15 regular students (qualified) for each teacher were selected at random from each class in the usually way. In few cases, intact classes were chosen.

¹. A list of schools visited is attached in Appendix- B (iii)
². Freeman, F.S., loc.cit., pp. 522-523
³. ibid.
4.30. Procedure for Administering the Final Test

Every care was taken before administering the test. The details of procedures were already discussed in Section 4.23 and therefore, repetition was avoided. However, one point had to be made explicit in this connection. In order to avoid 'halo' effect, which has been shown to be among the most serious causes of unreliability,¹ the students were instructed that the rating of each statement had to be made independently of other ratings.

The purpose of the study and the importance of conscientious response were clearly explained to the students both by the investigator and the Headmaster in few minutes talk followed by all sorts of assurance. The test was administered by the investigator himself with no one else in the room except selected pupils. The answered questionnaires were collected back as soon as they were found complete.

4.31. Methods of Scaling and Scoring

Usually, scaling of answers to a questionnaire having multiple choices involves giving "weights" to different response categories "in accordance with the degree of divergence from a typical answer."² Four different methods are found to be used. They are as follows:

1. Anastasi, Anne, loc.cit., pp. 144-145
2. Garrett, H.E., loc.cit., p. 323
(i) Thurstone method of "equal appearing interval",
(ii) Likert method of sigma (σ) scores,
(iii) Method of "empirical keying", and
(iv) Method of simplified scoring.

As our test (PAI) did not involve judgement of experts, the first method could not be used. The second method is applicable only when the distributions of response behaviours are normal. But as response-pattern was very complex (ref. Pre-tryout), the second method also could not be tried out. It was considered suitable to employ the method (iii). Strong in his VIB test and Jurgenson in his personality test also used this method for scoring purpose. The method of "Empirical Keying" is usually based on relative difference in frequencies of each response among two criterion groups. For the present test, the selection of criterion groups of teachers was made on the basis of Headmasters' Ratings on Teachers' Behaviour (HRTB) as per the plan of the work mentioned earlier (ref. 4:21). Rating for 45 (out of 49) teachers were obtained from their respective Headmasters. Assuming normality of distribution, about one third (14 out of 45) of the available teachers were chosen for our criterion groups (top 7 denominated as High Group and bottom 7 denominated as Low Group).

The answered questionnaires in respect of these two criterion groups of teachers were first separated. The responses

to each item were tallied according to four response categories: (a), (b), (c) and (d) for each group separately. The categorical responses for the two groups are shown in the Appendix-D(viii) and graphs are drawn for a comparable look (Ref. Chart 5).

By a careful examination of data in this table, it appears that scaling based on relative difference between two criterion groups is not logically sound so far as our sample is concerned. As for example, item 8 might be considered, in which, according to method (iii), the weight of 'a' (= always) appeared to be less than the weight of 'b' (= often) for teachers in the High Group thus leading to a meaningless conclusion. Similar things could be observed with items 4 and 12.

To overcome this difficulty, the fourth method ('simplified scoring' method) was tried out. Most studies of this sort indicated that a simpler weighing procedure would not substantially affect the reliability and validity of a test. ¹ The simplified process involves assigning plus one (+1) for a response having a positive weight and zero (0) for a response having zero weight or a negative weight. This is virtually the process of dichotomization as suggested by A.L. Edward. ² For our test, this method as advocated by Edward and Kilpatrik was employed because of its simplicity. We were justified to employ this method because the response categories used in our test were theoretically graded. The procedures followed in determining the weights to be given to response categories are discussed in Appendix-D(ix). The table for

---

CHART 5 Histograms Showing Responses in Different Response Categories of PAI Items by Pupils Under Two Criterion Groups of Teachers
(Data of Appendix D(viii) )
scoring keys is also presented in the said Appendix.

The significant difference in frequencies of dichotomised response alternatives among two criterion groups determines favourableness and unfavourableness. Plus one (or simply 1) indicates that the response occurred more frequently among high group than among low group. The zero (0) weight indicates that response occurred more frequently among low group than among high group.

The observed differences were tested for significance by constructing 2x2 contingency tables and then applying $X^2$ test by using the formula (71). The summary of $X^2$ values with probabilities is shown in Appendix-D(x).

The data in the table indicate that 13 items are found to be highly significant even beyond .01 level and items 1 and 12 are found to be significant at .05 level. This indirectly ensures the discriminative powers of individual items. It was, therefore, decided to use "simplified process" of weighing for scoring all the answered questionnaires by using score key Table (ref. Appendix-D(xi)) and following the method of scoring as discussed in Appendix-D(xii).

1. Garrett, H.E., op.cit., p. 265
4.32. The Consistency of PAI test Items - Final Form

Many survey studies indicated that the results might be altered by changing the test pattern in which questions and answers were expressed.\(^1\) As the final form of the present inventory suffered some changes, it was felt reasonable to review the consistency of individual items before proceeding further.

The internal consistency indices are essentially the measures of homogeneity of responses to "behaviour characteristics" expressed by the statements. Two methods, namely, "Empirical Keying" and correlation technique are usually employed for this purpose. The disadvantage of the latter technique was already discussed and so repetition was avoided.

Woodworth\(^2\) in his personal data sheet used arbitrarily selected criterion to differentiate normal from abnormal which, therefore, failed to demand statistical reliability of the measurement. Such procedure could not be accepted. Again Taylor\(^3\) and many others in the similar field used "exterior criterion" groups for measuring "internal consistency" of items. But this technique was not found feasible in our case. Since our sample scores were not normally distributed (details were discussed in Chapter 5), it was felt justified to use empirically selected criterion groups of teachers for measuring

\(^1\) Anastasi, Anne, loc.cit., p. 545  
\(^2\) Anastasi, Anne, chap. 18  
\( \phi_c \) co-efficients of items. The data in Appendix-D(x) could profitably be utilised for this purpose. \( \phi_c \) co-efficients were calculated by using the formula:

\[
X^2 = N \phi^2_c.
\]

Table 12: Showing Summary of Discriminative Power Indices (\( \phi_c \)) of PAI Test Items - Final Form

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>E</th>
<th>M</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.21</td>
<td>.51</td>
<td>.52</td>
<td>.43</td>
</tr>
<tr>
<td>2</td>
<td>.32</td>
<td>.63</td>
<td>.66</td>
<td>.22</td>
</tr>
<tr>
<td>3</td>
<td>.33</td>
<td>.36</td>
<td>.24</td>
<td>.18</td>
</tr>
<tr>
<td>4</td>
<td>.56</td>
<td>.30</td>
<td>.58</td>
<td></td>
</tr>
</tbody>
</table>

Though all \( \phi_c \) co-efficients were found to be significant (ref. Table 12) yet the magnitude of few, such as, for items 1, 8, 11 and 12 are found to be comparatively low. This might be due to the reason that our criterion groups were not sharply contrasted groups. In fact, it could be observed by a careful examination of data on HRTB that our contrasted groups consisted of high group and middle group (instead of low group) when considered in terms of percentage of score points. This could not be avoided. However, the values of \( \phi_c \) co-efficients were not so low as to render our instrument unreliable.

4.33. The Variability of PAI Ratings

As our chief aim was to measure effectiveness of classroom teaching activities mainly based on pupils' judgement, it was felt essential to examine first the degree of consensus of opinion (judgement) as well as dependability of judgement values. Usually, method of "averaging" is employed in such cases with considerably small degree of variability, though arbitrarily chosen. The variability is generally measured by standard deviation of scores when greatest stability is sought; and by quartile deviation when S.D. is influenced by extreme scores. When S.D. is very large, the dispersions of judgement values will be on a wider range indicating that respondents do not agree in rating the concerned teacher. On the other side, when S.D. is quite small, there will be a close agreement among respondents about a "typical score" point. So, the measure of dispersion is considered as an objective criterion of dependable ratings.

By a careful inspection of Master Data of PAI Test Scores (ref. Appendix-E(i) ) it would be appeared that distribution of total favourable responses assigned by the students to the teacher varied within a certain range. It was, therefore, essential to examine the variability of rating scores for each of the 49 cases separately to ensure the dependability of pupils' judgement.

1. Freeman, F.S., op.cit., p. 522
2. Garrett, H.E., op.cit., p. 60
CHART 6 Ogives (2 cases) of PAI-Scores Illustrating Extrapolation for Determining Q-Values (Data of Appendix E(ii))
It was apparent from data in Appendix-E(i) that S.D. would be appreciably influenced by the extreme score and, therefore, decided to use the "interquartile range" \( IQR = Q_3 - Q_1 \) as an index of variability. But as the sample of students for each class was small (only 10 in each class), a graphic method was employed to determine "Q-values" as close to their theoretical values. The cumulative frequencies for each distribution were computed and presented in Appendix-E(ii). The corresponding ogives were drawn by smoothing small irregularities in frequencies. The method of extrapolation as suggested by Gowda, was employed to those cases where a considerable agreement (in our case more than 25%) at any extreme interval was observed. The graphs are shown in Appendix-E(iii). But in order to illustrate this method, the graphs of two typical distributions (case Nos. 8 and 28) are discussed below:

In Fig 1 (Chart 6), the curve of case No. 8 does not cut the 25 percentage line and so \( Q_1 \) becomes indeterminate. In this case, to get \( 2Q_1 \), the difference in scores between \( Q_1 \) and medium is doubled and taken as 1.6 (= .8x2). In Fig 2 (Chart 6) the curve of case No. 28 is continued to meet the 75 percentage line and not the top extreme interval. The difference in score points between \( Q_3 \) and \( Q_1 \) gives \( 2Q_1 \).

The interquartile ranges for individual cases were noted in the respective graphs. It could be observed that these

1. \( 2Q = Q_3 - Q_1 \), where \( Q_1 = 25 \)th percentile (\( P_{25} \))
   \( Q_3 = 75 \)th percentile (\( P_{75} \))

CHART 7 Distribution Diagram of Means, Q-Values and Ranges of Ratings on PAI for 49 Teachers
(Data of Appendix E(iii))
values ranged from 1.2 to 4.4 with the average of 2.62.
But no empirical evidence from studies of similar nature was
obtained to substantiate our values. However, the $Q$ values
used by Thurstone and Chave in selecting statements for attitude
scale construction could be mentioned here in support of our
obtained values. They considered the judgement value of a
statement to be valid when its interquartile range was found to be
2.6. In support of this evidence, it could be said that our
observed values were not indicative of ambiguity or idiosyncrasy
in the ratings. It was observed that 57% of our cases ($N = 49$)
satisfied this prescribed value and 43% could not.

4.34. The Sex-Variability

In most studies of similar nature, it was indicated that
the boys and girls were not fundamentally homogeneous in their
attitudes towards teacher behaviour in the class. M.P. Dwivedi reported that the girls had a more appreciable attitude towards
their teachers than the boys. Kulandavel and Rao reported that
boys exhibited more heterogeneity in their ratings than the girls.

In view of these evidences, it was felt essential to
examine whether any sex difference was really indicated in the
present study. For this purpose, pupils of each class were
classified into two groups - one for boys and the other for girls.

---

1. loc. cit., p. 42
2. Dwivedi, M.P.
The average of total rating scores for each group was calculated. The differences in means (average) between these two groups were then tested for significance by using the formulae:

\[ t = \frac{m_1 - m_2}{\sigma_{m_1 - m_2}} \], where \( \sigma_{m_1 - m_2} = \sqrt{\frac{\sum d^2}{n(n-1)}} \)

Appendix-E(iv) presents the means of boys and girls for each of 49 teachers.

The value of 't' thus calculated is found to be 0.26 with df = 48. But to be significant at .05 level, the prescribed value of 't' is 2.01 for df = 48. The observed differences are, therefore, considered insignificant. So, in comparison with findings of the studies stated above, our findings showed a substantial improvement in achieving its objectivity in terms of sex sector variability.

4.35. Test Reliability, its Meaning and Methods of Estimation

"Reliability is the consistency with which a test yields the same result in measuring whatever it does measure". Usually reliability of a test is reported in terms of co-efficient of correlation called "reliability coefficient" between two sets of scores. Generally four methods, viz., test retest method,

1974, p. 429

1. Lindquist, E.F., op.cit., pp. 51, 58
equivalent form method, split-half method* and Kuder-Richardson method are found to be employed to measure reliability of individual test scores. But as our test intends to measure teaching effectiveness mainly based on pupils' judgement, it was felt justified to determine the reliability co-efficient of group judgement in addition to reliability co-efficient of individual test scores.

4.36. Reliability of Individual's Ratings

Method-1 (Test-retest reliability): This method was employed to measure the stability of the test. After a break of 10 days, the test was applied to a sample of students chosen from two schools at Kohima (CNT 1, 2, 9 and 10) who had already taken the test once. Altogether, 30 such students were available. Again after a period of about 3 months the test was applied to a sample of students from another two schools (CNT, 3, 4 and 7, 8) who had already taken the test once. Altogether 26 students were available. Table 13 shows two sets of paired scores at two different intervals. The scores on the first performance and the scores on the second performance of the same sample were correlated by using the formula (30)*. The reliability coefficient (r) at the first interval (after 10 days) is found to be 0.84 and at the second interval \( r = 0.73 \). They are obviously significant at .01 level. In order to compare our results with others, the


*Garrett, H.E., op.cit., p. 143, formula (30)
Table 13: Showing two sets of paired scores on PAI at two Different Intervals

<table>
<thead>
<tr>
<th></th>
<th>After 10 days ( n = 30 )</th>
<th>After 3 months ( n = 26 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

X = Scores on first occasion
Y = Scores on second occasion
average "r" was computed by Z-transformation\(^1\) and a value of average "r" = 0.80 was obtained. This value well agrees with the result of others in the similar field. Jayamma,\(^2\) in a test for teachers efficiency obtained "r" = 0.89 as a "test-retest" reliability. Our values of "r" for samples at different intervals thus indicated considerable stability.

Method-2: This method involves constructing of parallel forms in all respect and so, suffers from the disadvantage that twice as much material should be prepared as may be required for a single form. This method was, therefore, rejected.

Method-3: By a careful examination of data in Appendix-E(i), it could be observed that the degree of favourableness was different for different statements as well as for different classes. So, the conventional procedure of splitting up of items into two comparable halves could not be considered justified at this stage. A mathematically more defensive method for splitting up items into two comparable halves was employed. The average proportion of favourableness (here termed as popularity Index), based on the entire sample (N = 490) of students, was calculated for each individual item. The items were then arranged in rank order (high to low) of average Pi's and "odd-even" principle was applied in the usual procedure. The Appendix-E(v) presents the distribution of items in the two halves based

\(^1\) Garrett, H.E., op.cit., pp. 172-175

\(^2\) Jayamma, M.S., Construction and Standardisation of an Inventory for Predicting Teacher Efficiency", quoted in Buch's A Survey of Res. in Edn., 1974, p. 428
on rank order of average PI's of items.

In order to determine the reliability co-efficient based on those two equivalent halves, a selection of 50 students was made at random using random decimal digits of Interstate Commerce Commission\(^1\) following the conventional procedures. The aggregate scores of each student on the two halves were tabulated. Table 14 shows the pairs of scores on the two halves. The correlation between these two sets of scores was computed by using the formula (30)* and SEBF (79)** was used to get "r" for the pooled effect of the test as a whole and \(r = .82\) was obtained. The significance of \(r\) was determined by using the formula:\(^2\)

\[
t = \frac{r}{\sqrt{1-r^2}} \sqrt{N-2}
\]

The value of "t" is found to be 9.93 for df 48, \(P < .001\), showing that the obtained 'r' is highly significant. Our individual test scores are, therefore, highly dependable. Jayamma\(^3\) reported \(r = 0.86\) as "odd-even" reliability. Our results are corroborated with this result.

**Method-4 (Reliability Co-efficient by K-R formula "20")**: Methods of estimating the internal consistency of a test have been

---

1. Klugh, H.E., op.cit., p. 390
2. Lindquist, E., op.cit., formula (34), p. 211
3. loc.cit.

*Garrett, H.E., op.cit., p. 390

Table 14: Showing 50 paired Rating Scores on PAI for Two Equivalent Halves

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

X = Scores for first half
Y = Scores for second half
presented by Kuder and Richardson. In earlier section, no sophisticated statistical technique was used for this purpose. Such a technique was tried out at this stage. The sample \((N=50)\) which was used for estimating "odd-even" reliability co-efficient could profitably be utilised. The Appendix-E(vi) presents the favourableness and unfavourableness of each item made by 50 students. The values of "p" (= proportion of favourableness) and "q" (= 1 - p) are calculated. The aggregate score for each student is also given to facilitate computing SD of the total scores by using the formula (15). Finally "r" was estimated by applying the K.R.-formula "20":

\[
r_{11} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - \sum pq}{\sigma_t^2}
\]

and \(r (or \ r_{11}) = 0.763\) was obtained. This is obviously highly significant.

The summary of results obtained by different methods is given below:

<table>
<thead>
<tr>
<th>Test-retest</th>
<th>Split half ((n = 50))</th>
<th>K.R.-&quot;20&quot; ((n = 50))</th>
</tr>
</thead>
<tbody>
<tr>
<td>r (r)</td>
<td>.80 (average)</td>
<td>.82</td>
</tr>
<tr>
<td>(\sigma_r^<em>) (\sigma_r^</em>)</td>
<td>-</td>
<td>± .05</td>
</tr>
</tbody>
</table>

2. Garrett, H.E., op.cit., p. 53
3. ibid., p. 341

\[\sigma_r^* = \frac{1-r^2}{\sqrt{N}}\]
The data in the foregoing table indicate that the value of "r" by K.R. method is comparatively lower than the values obtained by other two methods. But the difference is so small as to render our test items inconsistent. However, on the whole, the table indicates a low value of r compared to those obtained by investigators in abroad for similar test although our results are well corroborated with the results obtained by others in Indian situation. Dr Brandon\(^1\) obtained "r" = .96 ± .01 and Cook and Carroll\(^2\) obtained "r" = .94. The reasons for obtaining relatively low value of r in our case are possibly due to (i) quality and maturity of students, (ii) small number of items, and (iii) variability of students. These could not be avoided.

4.37. Reliability Co-efficient of Group Ratings

The procedure that has been suggested by Lundberg\(^3\) in his "social Research", was employed in our case. The procedure involved selecting of two groups of 5 ratings each, chosen at random, for each of the 49 teachers and correlating average scores for the two groups and then stepping up the obtained co-efficient by S.B.P.F. as suggested by A.C. Rosander.\(^4\) Table 16 presents the average rating scores (rounded) for the two groups.

Table 16: Showing Average Rating Scores by Two Groups, 5 in each for 49 Teachers

<table>
<thead>
<tr>
<th>Cases*</th>
<th>Group I</th>
<th>Group II</th>
<th>Cases</th>
<th>Group I</th>
<th>Group II</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>8</td>
<td>26</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>9</td>
<td>27</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>11</td>
<td>28</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>12</td>
<td>29</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>10</td>
<td>30</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>8</td>
<td>31</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>5</td>
<td>32</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>33</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>11</td>
<td>34</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>10</td>
<td>35</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>7</td>
<td>36</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
<td>7</td>
<td>37</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>10</td>
<td>38</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
<td>9</td>
<td>39</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>10</td>
<td>40</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>3</td>
<td>41</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>17</td>
<td>9</td>
<td>10</td>
<td>42</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>18</td>
<td>8</td>
<td>7</td>
<td>43</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>7</td>
<td>44</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>8</td>
<td>45</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>21</td>
<td>8</td>
<td>4</td>
<td>46</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>12</td>
<td>11</td>
<td>47</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>23</td>
<td>8</td>
<td>8</td>
<td>48</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>4</td>
<td>49</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Case No. corresponds to code No. of teachers
Pearson's 'r' was computed by using the formula \((30)^*\) and 
\(r_1 = 0.702\) was obtained. The stepped-up estimate for this value 
was determined by using the formula \((79)^**\) and a value of 
\(r_u = 0.825\) was obtained. By using the similar procedure, Cook 
and Carroll\(^1\) obtained their "r" = 0.93 showing that our result 
well agrees with their result.

4.38. The Test Validity

"The validity concerns what the test measures and how 
well it does so." If the test fails to measure what it actually 
intends to measure, the purpose of the test would be deflated. 
The validity of the test was, therefore, considered very 
important, and it had to be measured with reference to the 
particular use for which it was intended.

According to A.P.A. Technical recommendations,\(^6\) there 
are four types of a test validity, designated as "content 
validity", "concurrent validity", "predictive validity" and the 
"construct validity". Though different procedures are employed 
to determine test validity of different types, fundamentally all 
procedures are practically concerned with the "relationship 
between performance on the test and other independent observable 
facts about the behaviour characteristics under consideration."

\*\* Garrett, H. E., loc.cit. p. 143
\*\* ibid. p. 339
1. loc.cit.
and diagnostic Technique", 1954, Vol. 51, No. 2, Pt. 2, 
quoted in "A Practical Introduction to Measurement and 
Evaluation" by Remmers-Rumpf Etc., p. 119
The content validity, sometimes known as "face validity" or "logical validity" is generally estimated by considering how well the test looks valid to subjects who took it. But according to Ebel the content validity largely depends upon the relevance of the individual's responses to test items rather than upon the apparent relevance of item contents. The predictive validity indicates the effectiveness of a test in predicting some future outcomes and is usually measured by correlating test scores with the corresponding measures of pupils' performance (objective criterion). The "concurrent validity" is more or less similar to "predictive validity".

For the present study the PAI test has been constructed on sound logic and scientific method. The teaching activities were conceptualised on the basis of important principles of effective mathematics teaching collected from various standard books on methodology and survey studies. Statements were selected empirically through successive stages, examined for ambiguity, consistency and reliability. The "inter-rater variability" was found considerably low and items were found well discriminated. The reliability was considerably high. Our test thus satisfied all the essential qualities of a valid test and, therefore, comparing the test results with any one of the standardised I.Q. test, achievement test or language test (as has been usually done) was not felt necessary.

However, it was considered useful to estimate the "concurrent validity" of the test. This was done by examining the agreement of test scores with corresponding class achievement scores in Mathematics (as criterion score). But as the school examination marks are non-standardised scores, the method of averaging of terminal examination marks was employed to minimise the probable error. Though the test was administered to 490 students, the class achievement scores in mathematics were available only for 360 students out of which 50 students were chosen at random in the usual procedure. The PAI test and corresponding CASM for each of these 50 students were presented in Table 17. A Scattergram was prepared as shown in Chart 8. The two regression lines were drawn. The closeness of these two lines indicated positive linear correlation between the two variables. The correlation co-efficient ('r') was calculated by applying the formula (30)* to the data of the Table 17 and $r = 0.67$ was obtained. It was tested for significance by using the formula (34)** and $t = 6.26$, df = 48, $P < .001$ was obtained. The validity co-efficient is highly significant.

This result well agrees with the results obtained by other investigators for similar tests. Jayamma¹ obtained $r = 0.66$ as co-efficient of correlation between marks of terminal examination of pupil-teachers and scores on her

* Garrett, H.E., op. cit. p. 143
** Lindquist, E.F., op. cit. p. 211

¹ loc. cit.
Table 17: Showing 50 paired scores on PAI(X) and CASM(Y)

<table>
<thead>
<tr>
<th>Case* Nos.</th>
<th>X</th>
<th>Y</th>
<th>Case* Nos.</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>9</td>
<td>47</td>
<td>342</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>84</td>
<td>12</td>
<td>59</td>
<td>346</td>
<td>8</td>
<td>49</td>
</tr>
<tr>
<td>89</td>
<td>12</td>
<td>60</td>
<td>351</td>
<td>12</td>
<td>38</td>
</tr>
<tr>
<td>136</td>
<td>8</td>
<td>25</td>
<td>355</td>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>141</td>
<td>11</td>
<td>48</td>
<td>358</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>144</td>
<td>8</td>
<td>30</td>
<td>359</td>
<td>12</td>
<td>47</td>
</tr>
<tr>
<td>148</td>
<td>9</td>
<td>26</td>
<td>361</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>150</td>
<td>12</td>
<td>22</td>
<td>362</td>
<td>10</td>
<td>46</td>
</tr>
<tr>
<td>151</td>
<td>4</td>
<td>12</td>
<td>381</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>158</td>
<td>3</td>
<td>14</td>
<td>385</td>
<td>12</td>
<td>60</td>
</tr>
<tr>
<td>160</td>
<td>2</td>
<td>13</td>
<td>390</td>
<td>12</td>
<td>54</td>
</tr>
<tr>
<td>174</td>
<td>9</td>
<td>32</td>
<td>391</td>
<td>6</td>
<td>47</td>
</tr>
<tr>
<td>175</td>
<td>5</td>
<td>44</td>
<td>400</td>
<td>9</td>
<td>49</td>
</tr>
<tr>
<td>180</td>
<td>7</td>
<td>44</td>
<td>401</td>
<td>9</td>
<td>52</td>
</tr>
<tr>
<td>185</td>
<td>14</td>
<td>74</td>
<td>406</td>
<td>5</td>
<td>38</td>
</tr>
<tr>
<td>187</td>
<td>5</td>
<td>10</td>
<td>411</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>292</td>
<td>7</td>
<td>23</td>
<td>416</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>295</td>
<td>12</td>
<td>43</td>
<td>419</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>298</td>
<td>9</td>
<td>15</td>
<td>428</td>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>300</td>
<td>14</td>
<td>41</td>
<td>442</td>
<td>9</td>
<td>27</td>
</tr>
<tr>
<td>301</td>
<td>3</td>
<td>17</td>
<td>445</td>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>305</td>
<td>2</td>
<td>20</td>
<td>450</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>341</td>
<td>10</td>
<td>34</td>
<td>457</td>
<td>10</td>
<td>25</td>
</tr>
</tbody>
</table>

*Case Nos. = code numbers of students selected at random from random digit table.
CHART 8 Scattergram and Correlation Table for 50 Paired PAI-Scores and CASM (Data of Table 17)
inventory. Samant Roy,\textsuperscript{1} in his study of teacher attitude obtained $r = 0.57$ as validity co-efficient for teacher attitude scale. On the basis of these evidences our instrument could be considered substantially valid for measuring classroom teaching in Mathematics.

\textsuperscript{1} Samant Roy, G.K., A Study of Teacher Attitude and Its Relationship with Teaching Efficiency, quoted in Buch's "A Survey of Res. in Edn.", 1974, p. 434