CHAPTER 2
A Review of Related Literature and Studies

2.1 Prologue:

Review is an activity which assumes a lot of importance in research. It is so because it provides the researcher with the information about the researches done in the past in the related area. This information enriches the knowledge of the researcher and provides direction for identifying the problem for his/her research. Since the review of related studies provides information about the research work already done and ultimately one gets the information about the research work which has not yet been done. Thus it gives insight into the problem and also the direction for dealing with the problem. The review of related literature provides knowledge about the literature available, the selection of appropriate tools that can be used and also the methodology of research. The following section deals with the literature and studies relevant to the present study.

2.2 Review of related literature

In order to have a better understanding of the problem under study, its methodological concepts and to select the appropriate statistical method and tools for analysing the data a thorough study of the related literature was undertaken.

2.2.1 About the Battery of Selection Tests:

The multiple choice objective-type test battery used for selection of employees in banks in the clerical cadre consisted of the following tests.
Test I: Reasoning Ability - This test consisted of 50 objective type questions. Each question had five alternatives out of which one was the right answer. Out of the 50 questions 25 questions were on verbal reasoning in which a problem was described and the candidates were supposed to solve it and 25 questions were of non-verbal type i.e. figures (diagrams) were given and the candidates had to find the answer by establishing the relationship between the given figures (diagrams). This test aimed at assessing the Analytical Ability of the candidate.

Test II: English Language - This test consisted of 50 objective type questions, each question had five alternatives out of which one was the right answer. Questions were based on a passage to assess the reading comprehension, functional grammar and lexical skills. The test did not require any knowledge about English literature. This test was aimed at assessing the facility with language one has. However, since it was a language test and had all the chances of being biased in favour of candidates from urban areas this test was only a qualifying test and scores on this test were not reckoned for merit-ranking.

Test III: Numerical Ability - Consisted of 50 objective type questions with five alternatives, one of the alternatives being the right answer. In this test 25 questions were on only computation while 25 questions were problems on arithmetic reasoning. This test assessed the proficiency in Maths. This test was aimed at assessing the facility with numbers and basic computation skill and did not require in-depth knowledge of mathematical concepts and formula. Besides computation, the 25 questions on arithmetic reasoning also assess the ability to interpret the problem described and solving it.
Test IV: Clerical Aptitude - Consisted of 50 objective type questions with five alternatives one being correct answer. 35 questions were on matching names and numbers while 15 questions were on classification and simple table reading. This test was to assess how speedily and accurately one can work with numbers and letters.

2.2.2 Standardization of Tests

Testing is a process used for testing the knowledge of an individual. It is used in schools and colleges to test the subject knowledge learnt as a part of curriculum. Besides the school/college education, there are tests which are used for testing various abilities that an individual has. The tests used are basically of two types (i) Objective type and (ii) Essay type. In an essay type test, the examinees are asked to write their own answers to the given questions, whereas in an objective type test the examinees are asked to choose the correct answer from amongst the given alternatives. However, the objective type items sometimes are of match the column or completion type items. These can be one word item or short one sentence answer item. The one sentence type items can come under essay type short answer items as well.

The assessment of essay type questions is bound to be influenced by the personal views, knowledge and style of the examiner whereas, the assessment of objective type questions is not affected by change in assessor, once the right answer key is prepared. Usually the teacher made classroom tests for the use of school/college examinations to be given to the school/college students are informal and not standardized. In the school/college examinations the tests used are teacher made tests based on the syllabus and when the scoring scheme is given assessment
can be done by anybody. The score given by any assessor will always remain the same.

The tests can be classified as informal or standardized.

Technically, a standardized test is a systematic process for assigning numerical values to sample of behaviour, such that the same or equivalent items are administered to all test-takers using uniform directions and scoring methods. Standardization can be extended to include uniform methods of test interpretation derived through the development of norms for the group for which the test was designed, information on the consistency of test performance (reliability) and evidence that the test measures what it was designed to measure (validity).

2.2.2.1 Performance Criteria for Standardized Tests:

Criteria for the performance of standardized test can be classified into two categories: (i) requirements for obtaining accurate and appropriate scores; and (ii) requirements for making accurate and useful interpretations of the scores. Requirements for obtaining accurate and appropriate scores include matched level of tests to the students, clarity and comprehensiveness in the administration of instructions and logistics. Requirements associated with test interpretations include norms, reliability and validity.
2.2.2.2 **Norms:**

Norms permit an interpretation of a test taker's score to be made relative to the scores made by a large number of similar individuals. These interpretations do not address how much content was mastered but how well the test-taker did in comparison to his/her peers.

2.2.2.3 **Administration of the tests under standard conditions:**

The conditions under which the tests are administered, play an important role in deciding the performance of the test-takers. It is therefore, extremely important and necessary to conduct the tests under specific conditions to have the valid interpretation of the test scores. Therefore for the standardized tests, not only the question paper and the answer paper are important but also the manual of instructions for test administration are equally important. The manual of instruction includes the following information.

(a) Description about testing conditions.
(b) Procedure of test administration step-by-step in detail.
(c) Directions to be given to the examinees by the examiner. The instructions which are to be read aloud are highlighted.
(d) Information required for answering general queries of the examinees.
(e) Time limits and the method to be followed for observing these time limits.
(f) Instructions for handling, packing and despatch of the test material.
It may be noted that the meaningfulness of the test results depends on strict adherence to the procedures described in the instruction manual (Wardrop, 1976, pp.2-3).

2.2.2.4 Objective scoring of the standardized tests:

Objective scoring means any two equally trained scorers would get exactly identical scores for the same test-paper. This is possible only for the objective type tests for which the scoring key can be fixed and the scorer has to just match the right answer key with the examinee's responses and count the responses which are correctly answered i.e. matching with the right answer key. Therefore the essay type of subjective tests are not used in standardized tests but only the objective type tests can be standardized. The scoring of a standardized objective type tests is not influenced by the subjective reactions of the scorers. The scoring of objective tests can now be done by using the automated scoring machine known as "Optical Mark Reader" (Wardrop, 1976, p.3.).

2.2.2.5 Construction of a Standardized Test:

The development of a test involves the following steps and aspects for consideration as described in Wardrop (1976, pp.72, 73).

**Test Content:** Content analysis is an extremely important aspect in development of a standardized test. Since the standardized test is expected to be used for many years, it attains greater importance to analyse the contents carefully. For content analysis the author is required to analyse the curriculum. Therefore not only the content analysis but also the curriculum analysis is required to be done. The curriculum should be
decided in such a way that it remains useful for the years to come. In the content analysis one has to prepare a **blue print**, which gives the topics to be covered, the number of questions from each topic and the distribution of difficulty level of questions under each topic and the weightage to be given for each topic. In short, the blue print of a test gives the detailed picture of the test.

**Item writing:** The next step to content analysis is item writing. To write good items is a difficult task. The items written must go through the stages of review, modification and editing before they are put to actual use. Even with the careful process of review, a number of items are likely to be discarded after actual try out. So at least two to four times the number of items should be written than that of the actual requirement.

**Item Analysis:**

Item Analysis is a technique to know the qualities of items, on the basis of responses given by the candidates who have taken the test. It is necessary to ensure the effectiveness of the test as a whole, in terms of each of its units i.e. items.

To contribute to make the test valid in terms of effectiveness of the test to tap the desired abilities, aptitudes, etc., there are certain features within the test. For a test to be valid, there is a pre-requisite that the test should consist of items which are effective and are free from any defect. Each test item should be homogeneous to the entire test, thus contributing to the purpose for which it is developed. A thorough item analysis also includes a number of quantitative procedures. Specifically, the numerical indicator are often derived during an item analysis: **item difficulty, item discrimination and distractor power** statistics.
Item Difficulty Index (p):

Item difficulty statistics is an appropriate choice for aptitude tests when the items are scored dichotomously i.e. correct vs. incorrect. The facility Index (p-Value) is the percentage of students who have given response as that particular answer choice. The p-Value of answer choice which is the 'key' for the item, is the percentage of students who have attempted the item correctly. Higher the p-Value easier is the item e.g. if the p-Value of an item is .68 means 68% examinees have answered it correctly, while if the p-Value of another item is .15 i.e. 15% of the examinees have answered it correctly. Thus the item having p-Value of .68 is easier than the item with p-Value of .15. It is expected for obvious reasons that more number of high scoring examinees answered the item correctly compared to the low scoring examinees. When p is the percentage of passing the item and q is the percentage of failing, pq is the variance. For p = .5, q = .5, variance = .25 which is the maximum possible value of variance. Therefore items with p-value .5 are most preferred items. Since .5 is the ideal value, practically the items with p-values ranging between moderate values of .4 to .6 are preferred to the much easier or much harder questions.

Item Discrimination Index (D):

The item discrimination analysis addresses the validity of the items on a test, that is the extent to which the items trap the attribute they were intended to assess. Similar to the Item Difficulty Index, the Item Discrimination Index can be derived whenever the items can be scored dichotomously as correct or incorrect.
Index of Discrimination is the item-test "point biserial correlation coefficient". The 'test score' of a candidate for a 50 items test is a continuum ranging from 0-50, whereas the 'item score' is a dichotomous score viz '0' if the response is not the correct answer and '1' if the response given is the correct answer. This dichotomous score of '0' or '1' is correlated with continuum score on the test. The statistical method used for this purpose is known as "point Bi-Serial Correlation" according to Garrett (1981. p.97).

When the items are scored dichotomously, the assumption of normality is unwarranted and therefore point biserial r is more appropriate than the biserial r which assumes normality of distribution. Moreover, biserial 'r' estimates the correlation on higher side as compared to point biserial 'r'. The biserial correlation is used when the dichotomous variable is a "forced" one like pass or fail and the point biserial correlation is used when the dichotomous variable is a "true" one like score of 1 or 0. In case of Objective test items the dichotomous variable is a "true" one which is score 1 or 0 and therefore point biserial correlation is calculated.

The index of discrimination can also be defined as the difference between p-values of high performing group (upper group U) and the low performing group (lower group L), and is designated as 'D'. Thus if the percentage of candidates in the upper group responding correctly to an item is 'U' and the percentage of candidates in the lower group responding correctly to the same item is 'L' then for that particular item D = U - L. Generally, items with .20 index of discrimination are acceptable while items with moderate index of discrimination above .30 are preferred.
2.2.3. Reliability

Reliability refers to the consistency of measurement. We would always like to use the tools that will give us a consistent measure of one's ability or aptitude; meaning thereby, the scores on a test taken at one time should not significantly differ from the scores on the same or similar test taken at some other time. But since the scores on psychological tests are derived from human responses, and one characteristic of human being is that their behaviour tends to fluctuate from time to time and from situation to situation.

**True Score** - Refers to the score that an individual would have obtained on a test under perfect conditions and it could be measured without error. However, this is a difficult, rather only a hypothetical situation so far as the psychological tests are concerned. Therefore, the scores obtained on a test can be represented by the following equation -

\[ X_t = X_T + E \]

where \( X_t \) = obtained score
\( X_T \) = True Score
\( E \) = Error which can be positive or negative.

It may be noted here that certain concepts viz. true score, error, true variance, error variance, etc. cannot be directly measured but they can only be estimated.

A score obtained on a particular test on a particular day should not be considered as an exact point but rather it should be thought of as
representing a zone in which the student’s true ability lies. That is, if a particular student scores at the 60th percentile, it is wrong to conclude that this is his exact standing. It is more appropriate to say that his true score lies somewhat near the 60th percentile. Scores obtained on any particular test on any particular day are somewhat biased. High scores are too high and low scores are too low. It’s not only the ability of the student that determines the score, but it is also the luck factor that contributes. The students who scored high on test, are not only high on ability but also they had a better luck at that particular point of time and the students who scored low are not only low on ability but also they had a bad luck at that particular point of time. If an alternate form of test is administered on some other day, the students who scored high on the first occasion, will as a group, score little lower while the group of students who scored low will as a group, score on little higher side. This is called the tendency of regression towards mean. However, this tendency does not necessarily apply to each individual. Few students who score high on the first occasion may score still higher on the second occasion and few students who scored low on first occasion may score still lower on the second occasion. All scores above mean tend to be somewhat biased upward i.e. they are probably higher than they should be.

2.2.3.1 Test-Retest Reliability

In this method of determining the reliability of test scores, an identical test is administered on a second occasion. The correlation coefficient between the test scores obtained on the first and the second occasion determines the reliability of the test. The error variance
corresponds to the random fluctuations from one test session to another. This could be due to any one or more reasons like change in testing conditions such as change in weather, light, venue, test administration personnel, distractions like noise, etc. or even a broken pencil point are some of the factors that may affect an individual's score. "Retest reliability shows the extent to which scores on a test can be generalised over different occasions. The higher the reliability, the less susceptible the scores are to the random daily changes in the condition of the test takers or of the testing environments.

While reporting the Retest reliability, the length of interval on which the test was readministered should also be mentioned. It is desired that the interval should not be very long since a remarkable development may take place over a period of time which may rise or drop an individual's status. For example, additional qualification or job-experience acquired over a period of time may lead to improvement in an individual's performance, on the other hand in case of old persons or housewives the performance may go down over a period of time. Therefore, the interval is desired to be short particularly in case of young children, in whose case the progressive changes are fast and can take place in a very short period of time.

Although apparently simple and straightforward, the test - retest technique presents certain difficulties. Though the interval is kept short, persistent practice may bring in some improvement in an individual's performance. Also, if the interval is short an individual may be able to recall his responses on the earlier occasion which may affect his performance. Retesting with the identical test does not have control over these factors.
2.2.3.2 Alternate Form Reliability

This method of determining reliability takes care of the difficulties found in the test - retest reliability. In this method an individual is tested on two different but equivalent forms on two different occasions. The correlation between the scores on two forms represents the reliability coefficient, gives us the measure of two types of reliabilities viz. temporal stability or the reliability of test scores obtained on different occasions and consistency of response to different item samples. Thus alternate form reliability provides more useful measures for evaluating a test.

Alternate form reliability should also be reported along with the interval of time between the administration of two tests. If the two forms are administered one after the other in immediate succession, the correlation coefficient represents the reliability across two forms and not across two occasions, and the error variance represents fluctuation in performance on two sets of items and not on two different occasions.

In this type of reliability, while developing the alternate form it is necessary to ensure that the two forms i.e. the two sets of items are truly parallel. The two sets of tests may be constructed independently with same specifications. The number of items under each area should be the same, moreover the items should be of same difficulty level. Time allotted for both the tests should be same, instructions and the sample items if any, should also be equivalent.

In this method the constraint of practice effect gets reduced but not completely eliminated. The forms of test being parallel, contain the problems based on the same principles. Once the principle is understood
by an individual, it may not be difficult applying the same principle and solving a similar problem. Moreover, developing truly equivalent forms may not be always possible.

2.2.3.3 Split-Half Reliability

In this method the scores are obtained for each student in two parts viz. for two halves of a test. The test is split into two equal halves, no. of questions in each part being equal. This can be done by taking first half and next half questions, or by taking alternate questions in the two parts or by any other method so as to ensure that the two parts of the test are equivalent. The method of splitting would depend upon the structure of the test. Usually, in most of the tests, first half and second half would not be equivalent, owing to differences in nature and difficulty level of items as well as cumulative effect of warming up, practice, fatigue, boredom and any other factors varying progressively from the beginning to the end of the test. A procedure that is most suitable for most of the purposes is to find the scores on odd and even items of the test. Such a method of division yields most nearly equivalent half-scores. However, while dividing the test into two halves, if a group of items is based on common information, all those items in the group should be put in the same part.

The split-half reliability provides a measure of consistency with regard to content sampling. It obviously does not provide the temporal stability of the test since the test is given in a single session. This type of reliability coefficient is sometimes called a coefficient of internal consistency, since only a single administration of a single form is required. Once the scores on the two halves are obtained, correlation between the two sets of scores is obtained. This correlation coefficient represents the
reliability coefficient. Actually, this type of reliability coefficient gives the reliability of half the test, since the number of items in each part is half of the items in the original test. Other things being equal, the reliability of a test increases with its length. A longer test will be more reliable. With a longer sample of behaviour, we can arrive at a more adequate and consistent measure. The effect that lengthening or shortening a test will have on its coefficient can be estimated by means of the Spearman-Brown formula given in Anastasi (1997, pp.95-96).

2.2.3.4 Internal Consistency Reliability

It involves single administration of a single form. It is based upon the consistency of responses to all the items in the test. The inter-item consistency depends upon the (i) content sampling and (ii) heterogeneity of the sample. The inter-item consistency increases with the homogeneity of the test. For instance, if a test contains questions on Maths, Science and English as against another test containing question only on English will have lower inter-item consistency. Further a test of English covering areas of grammar, vocabulary and comprehension as against a test of English only covering grammar will have lower inter-item consistency. Content sampling will depend upon the criterion that the test is meant for predicting. A highly heterogeneous criteria cannot be tested by a highly homogeneous test. In such case it is desirable to construct several homogeneous tests to measure different aspects of heterogeneous criteria.

The most common procedure for finding inter-item consistency is developed by Kuder and Richardson. It involves the calculation of Test-Reliability coefficients based upon the method of rational equivalence.
The method of rational equivalence is an attempt to get an estimate of the reliability of a test, free from the objections raised in other methods. Two forms of a test are said to be parallel when the corresponding items in the two tests are interchangeable and the inter-item correlation is same for both the forms. The method of rational equivalence stresses the inter-item correlations and also the correlation of items with the whole test. Out of the various formulas derived the most widely used one is known as “Kuder-Richardson formula 20” (KR-20). A simple approximation of above formula KR-20 is KR-21 as per Guilford (1978, pp.427-429).

2.2.3.5 Significance of reliability coefficient

There is no hard and fast rule as to how high should be the reliability coefficient of a test to be reliable. It depends upon the type of the test and the purpose for which it is used. As stated in Guilford (1984, pp.388-389), for research purpose, lower reliabilities are acceptable as compared to the one for practical purposes of diagnosis and prediction. Sometimes, we can make best use of the test even with a reliability as low as 0.50. For some purposes, even a test of low reliability adds enough to prediction to justify its use, particularly when used in a battery along with other tests. According to Kaplan (1977, p.121), reliability estimates in the range of 0.70 and 0.80 are good enough for most purposes in basic research. Also, since KR21, correlation coefficient is an approximation of Pearson's product moment correlation, and referring to table Q in Guilford (1978, pp.531-532), a correlation coefficient as low as 0.081 for df = 1000 is highly significant i.e. significant at .01 level, it can be concluded that any reliability coefficient more than 0.5 will be highly significant if the sample size is more than 1000.
An index of homogeneity given us the estimate about internal consistency of the test, however, it does not give direct indication of the amount of variability or error that can be expected in an individual test score. The difference between the obtained score and the true score is the error score or $X_t = X_T + E$ where $X_t$ is the obtained score and $X_T$ the true score. $E$ is the error which can be positive or negative. Standard error of measurement is the standard deviation of the distribution of error scores and is computed by $S_m = S_x \sqrt{1 - r_{tt}}$

Where $S_m$ is the standard error of measurement, $S_x$ is the standard deviation of obtained scores and $r_{tt}$ is the reliability coefficient.

95% confidence limits can be obtained by $X_t \pm 1.96 S_m$ that means the true score can be more or less by 1.96 $S_m$ than the obtained score. 99% confidence limits can be obtained by $X_t \pm 2.58 S_m$. Confidence limits give us the range by which the obtained score can differ from the true score.

2.2.4. Validity:

Any test is always designed with some purpose and the test scores have some meaning only when they are related to some other variable. e.g. an achievement test is meant to know the academic performance of students in the classroom, an aptitude test is meant for knowing certain abilities of an individual which in turn will predict the behaviour of an individual in certain domain. A test will be useful when its scores can be correlated with some other variable and can be meaningfully interpreted. If
a test has perfect consistency but its scores cannot be correlated with any other variable, such a test will be of no use and hence will have no validity. Validity is a generic term and can be defined at various levels in various ways. The validity of a test can be determined from the answers to the questions like – How well does the test measure what it is supposed to measure?, What traits does it measure?, Does it really measure what it is supposed to measure? Does it supply the information that can be used in decision making? What interpretation can be given to the test scores? What percent of the variance in the test score is attributable to the variable that the test measures?

Validity can be determined by the proportion of the true variance that is relevant to the purposes of the testing i.e. the extent to which the scores are attributable to the variable that test measures. Thus, as explained in Brown (1970, pp.97-98) the validity of a test is defined either by (1) the extent to which the test measures the hypothesized underlying traits, construct or factor, or (2) the relationship between test scores and some extra-test criterion measure.

2.2.4.1 Content Validity:

The class room tests are used for assessing the knowledge acquired by the students through the class room teaching. These tests are syllabus based and the items included in the test are just a sample as it is not feasible to cover entire syllabus in the test. The teacher thus has to test the entire knowledge through these test items and it is therefore necessary that the tests are representative of the behavioural domain. Scores on the tests are used not as ends in themselves, but rather to make inferences about performance in the wider domain. The purpose of
class room examination is to provide an objective basis for making an inference about the students' knowledge of the material covered in the unit over which they are being examined. Because it is not possible to ask every question of the unit, a sample of possible item is selected and is given in the form of a test. On the basis of the student's performance on this sample of items, his degree of knowledge in the entire unit of subject matter can be inferred. To the extent that the items are a good sample of the universe our inferences will be valid; to the extent that any bias is introduced in the selection of items, the inferences will be in error, and the test will be invalidated.

Although the concept of content validity is usually associated with achievement testing, it can also be applied to other areas of psychological testing. The definition of the content validity as per Brown (1970, pp.135-138) states that the items on the test must be representative sample of the universe of possible content or behaviours. Representative sample means taking a sample that includes in due proportion or frequency every relevant or required characteristic of the whole.

The principal method of determining content validity involves establishing one-to-one correspondence between the test items and the universe of behaviour with the help of an expert judge. However, this method does not have any quantitative index or even a set of agreed upon qualitative categories, secondly owing to individual differences, different judges will arrive at different conclusions. To overcome this problem, a well specified definition of the content universe is called for.
2.2.4.2 Face Validity:

It is not the validity in technical sense. It does not refer to what it actually measures, but it refers to what it appears to be measuring. Face validity refer to whether the test looks like valid to the examinees, the administrative users and other untrained personnel. Thus face validity is determined by a superficial examination of the test by the test-taker and considers only obvious relevances. Content validity, on the other hand, is established by the thorough and systematic evaluation by a psychometrically sophisticated judge and considers both subtle and obvious aspects of relevance.

Face validity is important to the extent that the appearance of the test has an influence over the motivation of test takers. For example, if in a particular situation if the test does not have face validity i.e. it does not appear to be valid to the test-taker, the test taker may feel the test has no relevance with the decision to be taken and hence may not be motivated to write the test, thereby affecting his performance. Thus, face validity, may not guarantee accurate measurement, may be an important influence on test-taking motivation and thus on the validity of the obtained scores.

2.2.4.3 Construct Validity:

Construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations represent appropriate evidence for this validation. Construct validity is said to be appropriate when the test user wishes to infer the degree to which the individual possesses some
hypothesised trait or quality (construct) presumed to be reflected in the test performance. Thus, whenever a test is to be interpreted as a measure of some attribute or construct that people are expected to be possessing; construct validity becomes important.

The construct validity is classified into two types viz. convergent validity and discriminant validity. When a test or some other measure of the proposed trait correlates strongly with other instruments designed to measure the same trait, it is said to have convergent validity. According to Guilford (1978, p.436), when a test correlates very little or not at all with measures of other traits; it is said to be having discriminant validity. Factor Analysis is the most commonly used technique to establish construct validity.

2.2.4.4 Criterion Related Validity:

The most common use of tests, other than the class room tests is to predict the performance in the domain of behaviour. For example, achievement tests are used to place the students in various class sections, while aptitude tests are used for predicting performance on job. The variable that is predicted by the test is called criterion, and the validity thus obtained is called criterion-related validity. Since the test is used to predict the criterion, the criterion-related validity is sometimes referred to as predictive validity; and since it involves collection of empirical data on the relationship between test scores and the criterion measure, this type of validity is also referred to as empirical validity. The proper measure of a test's criterion related validity, and thus its usefulness, is an index of its relative contribution, over and above that of other measures and sources of information, to increased decision making accuracy. We
are not interested in the test scores per se, but are interested in the test because it predicts some important criterion behaviour.

As described in Anastasi (1997, p.119), for certain uses of psychological tests, concurrent prediction is the most appropriate type and can be justified in its own right. The logical distinction between predictive and concurrent validation is based not on time but on objectives of testing. Concurrent validation is relevant to tests employed for diagnosis of existing status, rather than prediction of future outcomes.

2.2.4.4.1 Validity Coefficient method of determining criterion-related validity:

One of the most commonly used methods of determining criterion-related validity is by way of 'Validity Coefficient' i.e. by correlating the test scores with criterion scores. The procedure involves following steps as explained in Brown (1970, pp.109-111).

(1) Selecting an appropriate group to serve as subjects in the study
(2) Administering the test to the designated group
(3) Applying the relevant treatment
(4) Collecting the criterion data and
(5) Correlating the test scores with criterion scores.

The correlation coefficient thus obtained is called validity coefficient. Usually, the type of correlation coefficient is Pearson's Product Moment Correlation. The pre-condition for Pearson's Product Moment correlation is to have the linear relationship between the two variables to
be correlated. Therefore, it is necessary that the test-scores and the criterion scores are linearly related to each other.

One way of interpreting validity coefficient is comparative, i.e. to choose a test with highest validity coefficient. Being more valid, it is expected to be more useful. Validity coefficient can also be interpreted in terms of percent variance. The percent variance accounted for is obtained by squaring the obtained correlation coefficient. Thus, if, validity coefficient \( r = 0.5 \), variance \( = 0.25 = 25\% \). We can say that 25% of the variance is shared by the two measures, or that 25 percent of the variance in the criterion measure is attributable to variations in predictor scores. The validity coefficient can also be interpreted as a measure of predictive efficiency i.e. as the ratio of the average criterion score made by the persons selected by a ‘test’, to the average criterion score made by same number of persons selected on the basis of criterion scores. The major advantage of correlational validity coefficient is the ability to predict the criterion scores. Given the test scores, the criterion score can be predicted by using the regression equation \( y' = a + bx \); where \( y' \) is the predicted criterion score, \( x \) is the test-score; \( a \) is a constant to correct for the difference between \( x \) and \( y \) and \( b \) is the regression constant.

2.2.4.4.2 Decision Making Accuracy:

Since decision making on the basis of test performance is the ultimate aim, accuracy of decision is of utmost importance to the decision-maker. Decision maker's performance can be determined by calculating the proportion of accurate decisions out of the total decisions made. When the psychological tests are used for decision-making, the proportion of correct decisions made gives the index of the effectiveness of the test.
Higher the proportion of correct decisions made, higher is its effectiveness and validity. To calculate the index of decision making accuracy the subjects can be classified into groups on the basis of their performance in test and also on the basis of criterion performance. They can be classified into two groups acceptable or above average performance and unacceptable or below average performance. Thus we get the subjects divided into four groups as follows:

<table>
<thead>
<tr>
<th>Criterion performance</th>
<th>Unacceptable (- ve, below average)</th>
<th>Acceptable (+ ve, above average)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Decision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acceptable (+ve, above average)</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Unacceptable (-ve, below average)</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

The index of decision making accuracy is the ratio of correct decisions made to the total decisions made:

$$P_{CTot} = \frac{B + C}{A+B+C+D} = \frac{B + C}{N} = \frac{\text{Total Hits}}{N}$$

$P_{CTot}$, the index of decision making accuracy is referred as an index of validity when the validity is defined as accuracy of decision making.
In some situations when the success or failure of those who were not selected is unimportant, the ratio of successful selected persons to the total selected becomes more appropriate.

\[ P_{CPos} = \frac{B}{A+B} \]

A and B are the selected candidates who are unsuccessful and successful respectively. Pcpos is the index of validity in terms of ratio of selected persons who are successful to the total persons selected as described in Brown (1970, pp.118-120).

2.2.4.4.3 Interpreting Criterion Related Validity or Decision Making Accuracy Data:

Following factor influence the validity indices calculated in terms of Decision making accuracy.

A. Sample:

It is important to select the sample which will well represent the population. Size of the sample also plays an important role. As the size of the sample increases, errors of measurement tend to counterbalance each other and the obtained results could be expected to be more balanced. Therefore, with a larger sample, the results can be more expected to be statistically significant.

B. Base Rates:

As described in Brown (1970, p.127) the base rate may be defined as the rate of occurrence of a phenomenon in an unselected population. Thus, the base rate would be the proportion of people who would be successful on a job or in an academic program if there were no selection. In practical situations the selection is always made using some or the
other method applying some criterion, whatsoever it may be. Thus, in validity studies, the base rates more appropriately refer to the rate of occurrence of the phenomenon in relatively unsystematically selected group.

B. Selection Ratio: According to Kaplan (1977, p. 180), selection ratio is defined as the percentage of applicants selected or admitted. If the predictor is positively related to the criterion, by being more and more selective, we can increase the probability that any person selected will be successful. The effect of selection ratio on selection efficiency has been presented in Taylor and Russell's table. As the selection ratio increases, selection efficiency decreases for any validity index.

2.2.5 Job Performance:

Job performance is an important factor, which deals with the quality and the quantity of work done by an employee posted for a certain job. It's not only the quality of the work or the quantity of the work done by an employee but it is the linkage between the two i.e. how quickly as well accurately one can perform the assigned job. To determine the job performance of any role incumbent it is important to identify the key performance areas (KPAs) and various dimensions of the job assigned. Key performance areas are the critical functions to be performed by the role incumbent over a given period of time. It is important to define these categories meaningfully. These functions should specify what the employee should be doing rather than what results are expected from him. The key performance areas i.e. important functions to be performed by incumbent would obviously differ for different organizations. KPAs can be obtained by an elaborate and extensive job description. However,
broad job-descriptions, may not be helpful in identifying the KPAs as they do not give the exact and clear picture of the functions involved in a job.

Identification of Performance Areas may involve following steps:

1. A small group consisting of 4 or 5 persons may be formed involving those who are presently performing the role and those who have performed this role in the past and are presently supervising this role. The group may list out all the tasks associated with the role and then classify them into meaningful categories.

2. Out of the entire list of all the tasks, the group has to identify the main activities being performed, what is the exact role in performing these activities, how much time is spent on each of these activities and rank them in the order of importance to determine the weightage for each activity. The group may also spell out the expectations from a role incumbent if his performance is to be rated as excellent.

3. These activities may be classified into meaningful categories of functions, getting thereby the KPAs.

4. Assign weightages to the activities as per their importance.

Actual work behaviour i.e. job-performance is likely to result from a combination of motivational forces, moderated by skills and abilities. Even, a well executed plan of job-performance is doomed to failure if workers do not have skill and potential for the job. Both, performance outcomes and job satisfaction may be affected by aforesaid factors and may be affecting each other. The model
presented in fig. 2.1 depicts a few internal and external motivating factors which could be determinants of job performance in industrial organizations as explained in Rao (1984, pp.32-35).

**Determinants of Job-Performance**

![Diagram of Determinants of Job-Performance]

2.2.5.1 **Measurement of Job-Performance:**

Performance of an employee is usually measured by way of performance rating scale filled up by the supervisor (superior) or the reporting authority under whom the employee is directly working. This being an important tasks, it is very important and necessary that the ratings being given by the rater are unbiased and are made objectively. To achieve this it is necessary that the rating scales are designed with utmost care giving minimal scope for any subjective element. Reporting authority on the basis of his/her observations about the employees performance fills up
the performance rating proforma. Judgement about an individual's performance is made on the basis of these ratings. The performance rating scale is usually a five-point or seven-point scale. The assessment may be made using categories like "excellent" or "outstanding", "good", "above average", "average", "below average", "poor" etc. If categories are used they need to be defined and points may be assigned to them e.g. outstanding = 7, Very good = 6, Above Average = 5, Average = 4, Below average = 3, Poor = 2 and Very poor = 1.

In one type of appraisal system, the appraisee is asked to write the self-appraisal report giving details of the activities performed during the year of appraisal, and submits the same to the reporting authority. The appraiser (reporting authority) has to study the report and give his ratings based on his own observations as well as on the basis of appraisee's self appraisal report. This is followed by a one-to-one discussion between the appraiser and appraisee in which the appraiser can share additional information with the appraiser. In other type of appraisal system, there is no provision for self appraisal but the reporting authority has to make his observations and give his ratings on the basis of his own observations.

Smith and Kendall (1963) have evolved a procedure for developing evaluative rating scales anchored by examples of expected behaviours. The format proposed for these rating scales is a series of continuous graphic rating scales, arranged vertically. Behavioural descriptions, exemplifying various degrees of each dimension, are printed beside the line at different heights according to their scale
positions as determined by judgements of those who are expected to use scales. The behavioural descriptions are intended as anchors to define the levels of the characteristic, and as operational definitions of the dimension being rated.

The anchors are defined in terms of behaviours expected to be shown by the ratee. Ratings are to be made by checking at any position along the line. This type of rating scales are also called descriptive rating scales. Depending upon the requirement the proforma is descriptive or otherwise. In the descriptive rating scale the levels of performance are described and therefore it becomes easier for the rater to give his/her observations. The traits on which the employees are to be observed is decided on the basis of job-requirement situations. This differs from organization-to-organization depending upon the activities of the organization. The performance rating proforma designed for the use in present study is a combination of descriptive and objective rating scale. Traits decided and the levels described are decided taking into consideration the job requirements in banks at clerical level.

2.2.6 Job-Satisfaction:

According to Vroom (1978, p.99), the terms job-satisfaction and job-attitude are typically used interchangeably. Both refer to effective orientation on the part of individuals toward work roles, which they are presently occupying. Positive attitude towards the job is conceptually equivalent to job-satisfaction and negative attitudes towards the job are equivalent to job dissatisfaction. Job satisfaction plays a central role in the study of behaviour at work.
Knowledge of the determinants, the consequences, and other correlates of job satisfaction are vital. Once an individual joins an organization, a vector of scores on a well-constructed, validated set of job satisfaction scales becomes the most informative data. Job satisfaction is an emotional reaction to a job that results from the incumbent's comparison of actual outcomes with those that are desired. General job satisfaction involves components not caused by the immediate job-situation. One is temperamental or happiness, another is trust in management. Both can act as causes, effects or quasi moderators, and each is likely to be related to co-operative and adaptive behaviour. Since neither can be changed easily, both should be measured and the extent of their influences estimated. The management of an organization may considered changing certain aspects of a job-situation, viz. change in job-design to improve satisfaction with task characteristics, or introducing training to improve supervisor's interpersonal skills. Changes are generally introduced in the hope that improvements in facet satisfaction will in turn affect broader areas of employee satisfaction and eventually improve behaviour and reduce costs. The managements in general are interested in improving job satisfaction. Greater the job-satisfaction, better is the quality of life, health and stability. Following points have an important place in determining the job satisfaction as explained in Cranny et.al (1992, pp.45-47).

1. Job-satisfaction must be measured reliably.
2. Different results may be expected according to the scales used.
Satisfaction cannot be judged in absolute terms, but involves comparison.
The relationship between person and environment is interactive.
Job satisfaction can be broken down into facets or components.
Each of these facets can be tied to one or more aspects of work environment and the job.
Satisfactions with these aspects are inter-correlated defining a general overall factor.
Some of this general factor can be attributed to relatively permanent characteristics of the individual.
Job level is typically correlated with satisfaction with all aspects of the job.
Community characteristics can account for a large proportion of the inter-correlation of satisfactions when the sample has been drawn from a number of communities.

Job satisfaction, as observed by Herzberg consists of two distinct dimensions - job-satisfaction and job-dissatisfaction. These are two different states and depend upon factors - hygiene factors and motivational factors. The hygiene factors viz. working conditions, salary, inter-personal relationship etc. when present do not lead to job-satisfaction but when not present they bring in job-dissatisfaction. Hence they are the dissatisfiers. The motivational factors viz. achievement, responsibility etc. when present lead to job-satisfaction. Therefore, satisfaction depends upon motivational factors while dissatisfaction results from the absence of hygiene factors.
The job-satisfaction is to be treated as a complex set of variables, because the workers who are highly satisfied with one dimension of their job e.g. work-itself, may be dissatisfied with another dimension say "wages". It is therefore theoretically, as well as practically useful to consider it as a set of different dimensions.

2.2.6.1 Measurement of Job-Satisfaction:

Job-satisfaction of employees is usually measured by way of questionnaire. The questionnaire consists of questions aiming to know about the employee's feeling about the work assigned and being performed by him/her. The questionnaire is given to the employee for filling up and indicating his/her feelings about the job. Another way of measuring job-satisfaction is interviewing the employee and asking him/her the feelings about his/her job. Job-Description-Index (JDI) is one such questionnaire designed by Smith, Kendall and Hulin. It gathers information about employees feelings about the work itself, supervision, people (Co-workers), pay and promotion. The researcher has used JDI as tool for the present study.

2.2.7. Multiple Regression Analysis:

Multiple regression is a technique used to predict scores on a single outcome variable Y on the basis of scores on several predictor variables, the XiS. Multiple correlation is the estimate of the correlation between one dependent variable and two or more independent variables. When the number of independent variables becomes more than two, the calculations become too tedious to be carried out manually. Therefore, the calculations are required to be
carried out mechanically using software packages. It gives the estimate of proportion of variance in the dependent variable that is accounted for by the independent variables.

The multiple regression equation is used for prediction of a dependent variable from the knowledge of values of independent variables. The prediction equation for two independent variables $X_1$ & $X_2$ and one dependent variable $Y$ is $Y = b_0 + b_1 X_1 + b_2 X_2$ as mentioned in Harris, (1985, pp.50-51) where $b_1$ and $b_2$ are weights and $b_0$ is the constant. The equation gets expanded with increase in number of variables.

2.2.8 Item Characteristic Curve (ICC)

A valuable way to learn about item is to graph their characteristics, which can be done with the 'Item Characteristic Curve' (ICC). It is a curve plotted by taking test scores along X-axis and the p-value or Item-facility index along the Y-axis. It is expected of a good test-item that more number of high performing candidates answer the item correctly as compared to the low performing candidates. Therefore, it is expected that, as the test-score increases the p-value also increases and the ICC for a good test item should be rising up as we move from left to right on the X-axis. An ideal ICC is shown in Figure 2.2.

![ITEM CHARACTERISTIC CURVE](image)  

Figure 2.2
2.3 Review of related studies:

A large number of studies were reviewed with a view of getting insight into the problem. Of these, the following studies were found relevant to the present study.

2.3.1 Studies related to Job-Performance and Job-Satisfaction:

(A) Studies conducted by NIBM/IBPS

Deshpande, et.al (1974 pp.417-444) studied - Relationship Between Job Performance of Clerical Employees and Type of Selection Test: Objective and Descriptive.

Objectives of the study was "To compare the performance of two groups recruited at about the same time through two different - objective and descriptive - examinations". The two groups of candidates (i) recruited through NIBM objective type tests and (ii) recruited through traditional descriptive type test were selected as sample. Data as regards their performance was obtained from their supervisors through a mixed rating scale - performance rating proforma. The data were analysed for comparison on different abilities which are desired of a bank employee.

Findings of the study were - (1) The candidates selected on the basis of job-related objective tests proved to be not only effective job performers but they were also equally good or better performers than those recruited through the traditional type of descriptive examination. (2) The results of the pilot study also indicated that candidates recruited on the basis of objective tests possess higher
potential for promotability as compared to those recruited on the basis of a traditional examination.

The results of the study did indicate the strength of objective tests vis-à-vis the descriptive papers in clerical selection. However, it must be borne in mind that the tests have to be related to the important components of the job and also have to be carefully constructed. Any objective test poorly constructed or improperly administered may not prove effective.

Mankidy (1977, pp.64-79) carried out a study - How appropriate was the clerical selection?: A validation of NIBM clerical selection Tests.

Objectives of the study were (i) to study the validity of selection tests for clerical recruitment in banks or how far the selection had achieved the contemplated and desired objectives, (ii) To find out whether the selection tests had given to the banks "appropriate" type of clerical recruits.

The validity of the tests was tested against the criteria "on the job performance". Findings of the study were (1) A considerably large proportion of employees recruited through NIBM were evaluated by their supervisors as very good performers. Even in regard to the specific abilities, the evaluation was on the same lines. (2) It confirmed that the selection strategy evolved at NIBM had been more effective in selecting clerks with higher potentials for supervisory and management jobs in comparison with the traditional system of selection.
(B) Other Studies:

Amarsingh (1985, pp.1069-1070) studied, Correlates of Job Satisfaction Among Different Professionals.

The important findings of the study were (i) The job-intrinsic variables such as job-concrete and job-abstract correlated positively and significantly with job-satisfaction of professionals. (ii) The job-extrinsic variable including psycho-social, enconomic and community growth factors was found to be positively related to job-satisfaction of professionals. (iii) age was found to be positive correlate of job-satisfaction. (iv) Experience correlated positively and significantly in case of advocates and doctors with job-satisfaction while in case of teachers and engineers this correlation was not significant. (v) Self-esteem was found to be positively related with job-satisfaction. (vi) high scores on extraversion affected the job-satisfaction of teachers, engineers, advocates and doctors negatively.

Chaudhari and Lahiri (1968, pp.41-62) - Carried out a study on Perceived Job Characteristic as Satisfier and Dissatisfier by Manual Workers.

The study was based on the work attitudes of 100 skilled blue-collar workers. Questionnaires were employed to determine worker's satisfaction with 8 job-context factors and 5 job-content factors. Satisfaction index for each factor was derived by subtracting the importance score (need-strength) from the extent to which the
need was met (input). Satisfaction was predicted when the difference was equal to or greater than zero and dissatisfaction was predicted when the difference was less than zero. The results showed that hygienes as well as motivators may contribute to sources of worker's dissatisfaction. It was concluded that the two sets of work attitude determinants known as job-context and job-content factors were not independent of each other as sources of employees satisfaction and dissatisfaction at least for blue-collar workers.

Hulin (1968, pp.122-126) studied - Effects of changes in Job Satisfaction levels on employee turnover.

It was observed that the turnover rate of female employees of the company was as high as 30%, which was higher than other companies in the area. To find out the reason, the researcher administered JDI to the 345 female clerical workers. The results of the scores on JDI indicated that dissatisfaction of clerical workers was occurring in all areas. Analysis of their responses substantiated this finding. After the survey study and implementing changes in the company's policies, this rate came down. Moreover, there was a significant increase in satisfaction with four out of five of the job areas.

These findings indicated that the difference in the multiple correlation was non-significant. Even with a turnover rate of 12% a significant amount of the variance in individual termination decisions was attributable to differences in satisfaction.
Kolte & Supe (1972, pp.405-413) studied - Determinants of Job-satisfaction of village level workers: a test of herzberg's dual-factor theory.

The study was conducted in Vidarbha region of Maharashtra. The V.L.Ws were given the schedule in the training class. They were assured of the anonymity of their responses and rapport was built-up. The schedule was connected with their job and were asked to describe. The factors responsible for inducing job-satisfaction and job-dissatisfaction were presented. It was revealed by the critical incidents that satisfies, in order of the frequency of incidents in which they were cited, included achievement, recognition, work itself, good superior, advancement, money and responsibility whereas dissatisfiers included, again in the order of frequency, policy and administration, work itself, interpersonal relations, working conditions, advancement, salary, lack of recognition and poor technical supervision.

Thus the study explored the factors associated with both job-satisfaction and dissatisfaction and found that the factors were not necessarily distinct and separate. The factors were found to interact among themselves; the feelings of either satisfaction or dissatisfaction being determined by the achievement of the aim of promotion and salary. Thus, Herzberg's dual factor theory was not supported by the empirical evidence in this study.

Koustelios and Baoiatis (1997, pp.469-475) University of Thessalv studied - The employee Satisfaction Inventory (ESI): Development of a scale to measure satisfaction of Greek employees.
A pool of 130 items was collected by interviewing Greek employees. It was reduced to 83 items. These items were distributed over the six subscales viz. working conditions (12 items): Supervisor (24 items) pay (12 items): Job itself (17 items): Organisation as a whole (12 items) and Promotion (9 items).

The selection of items for inclusion in the final subscales was carried out and finally it got reduced to 24 items from 83 items representing the six refined subscales.

The subjects were requested to indicate their agreement with each item on a 5-point scale ranging from strongly agree (1) to strongly disagree (5).

The main purpose of this study was to develop an instrument to measure employee's job satisfaction based on Greek samples. In this respect, the factor analytic results were encouraging: a six factor solution emerged. Further, results from confirmatory factor analysis confirmed that the six factor model was fairly good.

Kulsum U. (1985, pp.1090-1097) studied the Influence of School and Teacher Variables on Job Satisfaction and Job Involvement of Secondary School Teachers in the City of Bangalore.

The final sample of the study had 586 secondary school teachers selected on a proportionate stratified random sampling technique. Tools used to measure the job-satisfaction and job-involvement inventory were Indiresan's job satisfaction inventory and a job involvement scale respectively. Winer's Leadership Behaviour
Description Questionnaire (LBDQ) and Lawler and Porter's Job Performance Scale were used to measure the leadership behaviour and job-performance respectively. To measure the attitude towards teaching profession and a scale on teacher effectiveness were constructed by the researcher. Sharma's SOCDQ was used to quantify organizational climate.

Findings of the study were - (i) Permanent teachers had a higher level of job-involvement as compared to temporary teachers, (ii) Teacher's job satisfaction, teacher effectiveness, teacher's attitude towards the teaching profession, students size and teacher's performance turned out to be the significant predictors of teacher's job involvement, (iii) Teacher's attitude towards the teaching profession and teacher effectiveness turned out to be the common predictors of both job-satisfaction and job-involvement.


The investigation was carried out in three textile mills of Ahmedabad, on 230 clerical and 170 supervisory technical personnel. Two short performance standard inventories, each having ten items were used for evaluating the performance of subjects in the sample. Seven out of 12 GATB were selected and administered to the same sample. Means and SDs of all the scores were computed separately for two groups across the different
demographic variables. ANOVA and t-test were applied to test the significance of differences in means.

The main findings were - In clerical group (i) there were no significant differences among the mean scores across departments, (ii) personnel with length of service up to 25 years showed significantly higher mean scores than those having length of service from 26 to 35 years, (iii) on some parts of GATB, graduates scored significantly higher than matriculates while matriculates scored significantly higher than non-matriculates.

Findings for supervisory technical group were also similar.

2.3.2 Studies Related to Validity and Reliability:


Sample consisted of 205 students from six high schools; with males and females in almost equal proportion. In order to test the validity of the three factor model of learning approaches, structural equation modelling techniques were employed. The present problem required the use of confirmatory factor analysis. The overall fit of the data to the model was very good resulting in a comparative fit index of 0.97 based on 11 degrees of freedom. As required for the test of the model, all three latent variables were maintained in an orthogonal configuration. The lack of correlation between the three latent variables and the resulting good fit of the
data to the model, support the theoretical orthogonality of the endogeneous factors.

The results of the study strongly supported Biggs three factors model underlying the Learning Process Questionnaire. The major results indicated that (1) the overall fit of the model to the data was very good. (2) The three basic factors were clearly identified and were orthogonal and (3) particular parts of the model did not have a good fit.

Gokhale (1981, pp.149-153) studied Validity of tests used in the selection of spearhead teams.

The sample consisted of 28 persons who were finally selected for the job. They were in the age-group of 21-28 years and were from rural background. At the end of six months training in rural background they were rated on a Three-point scale on Six dimensions. The test-retest reliability of the scale was 0.8. Correlation coefficients between these six dimensions and the dimensions underlying tests were calculated.

It was found that the tests used were appropriate for predicting only some measures of job performance. In fact traits viz. hard-work, convincing people and planning had very poor values. The test battery, however, needed to consist of tests which could be sequentially given so that the desirable traits could be held common and other cognitive tests then administered to assess the planning and other abilities of the applicants.
Puhan (1978, pp.95-100) studied, Psychometric Invariance in Reliability and Validity Contexts. Psychometric invariance as a concept and as a requirement of psychological tests, was relatively new. Its implications, usability, and overall suitability as a scientific concept in psychometrics, would largely depend on its meaningful relations with other existing requirements of psychological tests, viz. reliability and validity. Therefore, the basic purpose of the study was to facilitate logical connections between the concept of psychometric invariance and the concept of reliability and validity. It could be concluded that psychometric invariance of a test were far more useful than those of traditional reliability and validity assessment of a test. A psychometric invariance assessment across time would examine the stability of factor loading patterns of that test across two different occasions. This would represent a case of test-retest reliability. Such an invariance assessment could also be used to test the validity. Psychometric invariance assessment simultaneously evaluated reliability and validity of a test in terms of its factor loading components.

2.3.3. Studies Related to Construction and Standardization of Tests:

The sample consisted of 5,449 students selected from ninety-six different schools of sixty-two different places of Saurashtra region by using the technique of stratified random sampling. Initially the
test consisted of 200 items for pre-tryout. After the item-analysis, 134 items were retained for the pilot test and divided into two forms. The final form of the test consisted of sixty items.

Descriptive statistics like central tendencies, SD and Skewness were worked out. Percentile scores, standard scores, T-Scores and Stanines were developed. Reliability was established by test-retest, split-half and Kuder-Richardson formulas 20 and 21. The reliability coefficients obtained by these four methods were found to be 0.82, 0.93, 0.91 and 0.82 respectively. Validity of the test was established by correlation with intelligence tests, aptitude tests like abstract reasoning, numerical ability and verbal reasoning test.

Conclusion: (i) The means of boys and girls of Grade IX were higher than those of Grade VIII. (ii) The means of boys were higher than those of girls in Grades VIII and IX in the total sample and (iii) Urban and rural area differences were observed only in the case of the Grade IX sample.

Deb Maya (1966, pp.73-76) studied, Standardisation of a Group Intelligence Test Earlier days Binet Scale or its adaptations were commonly in use as intelligence tests. However, as these tests being of individual type could not be administered to a group at a time. Shri S K Bose and Shri S C Datta from Calcutta University, constructed a test comprising questions in the areas (1) Arithmetical Reasoning, (2) Verbal Reasoning, (3) Vocabulary, (4) Direction, (5) Number Completion and (6) Analogy. It was a speed test. Subsequently, the author of this paper undertook the standardisation of the test.
The test had an uniform procedure of administration and scoring. For each item there was only one predetermined correct answer. No personal equation of the administrator could alter the test-score. Thus the test could be considered as an objective test.

The scores were found to be normally distributed with the Mean score of 35 and SD of 8.7 having range of 10 to 59.

Reliability was tested by the split-half method. Correlation Coefficient with Spearman-Brown formula was found to be 0.92, hence highly reliable.

The test was earlier validated against academic achievement. The correlation was found to be 0.38. (significant at .01 level), which was statistically significant. Later on the test was validated against a standard intelligence test, Raven's Standard Progressive Matrices. The correlation was found to be 0.67 (significant at .01 level). Thus the test could be considered as valid.

Mishra,(1985, p.354) studied the Construction and Standardization of a Verbal Group Test of Intelligence in Oriya for the Age Group 12+ to 15+.

The item areas of the test were verbal analogy, verbal reasoning, vocabulary general information and numerical relations. The test was standardized on a sample of 2000 boys and girls chosen on a stratified random basis. Split-half, test-retest and other reliability coefficients were calculated. The inter-item correlation
and factor analysis with varimax rotation were used for a study of validity of the test.

The study resulted in developing a verbal group test of intelligence in two parallel forms. The test had 5 subtest areas of 50 items and required 30 minutes for administration in the classroom situation using answersheets.

The reliability indices were high in the range of 0.73 to 0.92; and the validity coefficients in the range of 0.52 to 0.73. The factors identified through factor analysis were general reasoning and verbal comprehension.

Pillai K K (1978, p.500) studied The construction and standardization of a Verbal Test of Intelligence in Tamil (for the age group 10 + to 15+)

For pre-tryout a sample of 100 students was used. For further try-out a sample of 750 students selected from three schools in Chidambaram was used. For final administration, 5000 students from 34 schools in one of the districts in Tamilnadu were selected by the method of stratified proportionate sampling. The test used included seven sub-tests: Synonym, antonym, analogy, classification, mixed words, reasoning (Verbal) and reasoning (numerical). A total of 110 items were there in the test.

The coefficient of reliability computed by test retest method was found to be 0.84 and that by the split half method was found to be 0.88. The content validity was considered on the basis of various
types of behaviours assessed by the sub-tests. Norms were determined in respect of the total sample, grades and age groups.

2.3.4 Study based on Factor Analysis:

Nair, K.S. (1972, p.493) carried out An Analytical Study of the Factor Pattern of Verbal and Non-Verbal Tests of Intelligence.

The sample for item analysis consisted of 370 students and the sample for final analysis consisted of 420 students chosen from the secondary schools of Trivandrum educational districts.

Correlations among the sixteen variables were computed using the Pearson's Product moment method of correlation.

(a) Verbal and Non-Verbal tests were formed mainly on the basis of content.

(b) A third factor, identified as numerical ability, was the same as the one identified by others.

(c) A fourth factor which showed possibilities for tests to be grouped on its basis, would have emerged if more tests had been used.

(d) Factor I which had high loadings on the tests of analogies, series, spatial relations, classifications, water-reflection and arithmetic reasoning could be identified as Non-Verbal factor.
(e) Factor II which had high loadings on vocabulary tests as well as water reflection which was classified as a Non-Verbal item could be termed as a Verbal factor.

(f) Factor III which had high loadings on arithmetic reasoning tests, number series and number classification, could be termed a numerical reasoning factor.

2.4. Implications of the Review:
A thorough review of the related studies and literature was of great help as it gave information about the studies already carried out by various researchers in this area.

2.4.1. Implications of Review of Literature:
While reviewing the literature on validity and reliability, details of various methods and their relevance in different situations was studied. It was observed that the product moment correlation for calculating validity coefficient was the most widely used method. The concept of 'Criterion Related Validity' as 'Decision Making Accuracy' was found to be most relevant for present study since most of important part of the selection process in banking industry is the decision making accuracy in selection of personnel based on the objective type tests. Therefore, it was decided to use this method for calculation of validity coefficient, though it was not used by the researchers of the studies under review. KR-21 formula was used for establishing the reliability (internal consistency) of the tests. Multiple Regression Analysis was used for estimating the Job-Performance and Job-Satisfaction on the basis of performance on objective type tests. Literature on Job-Performance was useful for designing the performance rating proforma for the present study, as it
gave information about various traits on which the performance should be rated in different situations and also about the types of scales that could be used in different situations. Literature on Job-Satisfaction gave information about how to measure job-satisfaction, its importance in job-situation, the elements considered while designing a tool for measurement of Job-Satisfaction and also the various tools which were readily available. Job Description Index i.e. JDI, a tool designed by Smith et.al. was chosen for the present study since found relevant for the situation.

2.4.2 Implications of Review of Studies:

2.4.2.1 Implications of Studies Related to Job Performance and Job-Satisfaction

(A) Implications of Studies conducted by NIBM/IBPS

The study of Deshpande et.al (1974) and Mankidy (1977) had given the information about the types of tests being used at the inception of the system, work done by NIBM in relation to testing as selection strategy for bank recruitment. It answered the question like why more importance was given to the objective type tests as compared to the 'traditional' descriptive type tests and how the selection system was evolved. Since the present study basically dealt with the present selection strategy for bank recruitment and was an effort to establish the validity of objective type tests, the above information was very much required and useful. Mankidy's study also threw light on how the validity of tests was established against the criteria of job-performance, tools and traits selected for measuring job performance. Present study was an extension
of these studies in the changed scenario. Review of these studies confirmed the fact that though there were changes in patterns of tests used by NIBM/IBPS, no such study was undertaken in the past more than 20 years.

(B) Implications of other Studies:

Charles Hulin, had used JDI for measurement of job satisfaction while studying reason for high turnover rate among female employees. This study told us how the job satisfaction or dissatisfaction affects an employee and areas of satisfaction covered by the JDI. Job satisfaction was one of the criteria used for establishing validity in the present study and the JDI was the tool used for measuring job satisfaction, Hulin’s study was useful in obtaining the information particularly about the effect of job satisfaction or dissatisfaction on an employee. Studies by Kolte & Supe (1972), Koustelios & Baoiatis (1997) and Amar Singh (1985) provided information about the factors which determine the job-satisfaction of an individual and the types of tools that could be devised and used for measuring the job-satisfaction of professionals in different areas at different levels. This helped in deciding the factors relevant for the present study and selecting the tool for measurement of job-satisfaction. The study by Kulsum (1985) dealt with the relationship between various factors and their effect on job-satisfaction and job-involvement of teachers from different strata. It involved measurement of job-satisfaction, job-performance and job-involvement and establishing inter-relationship among them. It helped by giving insight into how to measure various factors using different tools and how to establish the relationship; what were the factors affecting job-satisfaction and job-performance.
Study of Roy A (1982) dealt with finding out the factors that affect the job-performance. This threw light on what were the important traits on which the performance should be assessed and about the tools for measuring it. Study of Chaudhari & Lahiri (1968), dealt with the analysis of job elements that act as satisfiers and dissatisfiers for blue collared workers. Thus, it gave us information about the factors determining job-satisfaction of the employees.

Studies on job-satisfaction and job-performance provided information about the factors working as satisfiers and dissatisfiers methods and tools for measurement of job satisfaction and job-performance and development of scales for the purpose which was very much useful for the present study.

2.4.2.2 Implications of Studies Related to Validity and Reliability:

Gokhale (1981) studied the validity of tests with a purpose of predicting occupational success and finding out the utility of tests in predicting various traits of job performance. This study along with other studies on validity and reliability threw light on different methods of establishing reliability and validity. Since the purpose of the present study was to establish the validity and reliability of the tests and ultimate purpose of the tests was predicting occupational success, these studies were helpful.

2.4.2.3 Implications of Studies Related to Construction and Standardization of Tests:

Studies by Bhatt (1981), Pillai (1978), Mishra (1985) and Maya Deb (1966) deal with the construction and standardization of objective type tests in different areas as such Test of Reasoning (verbal and non-verbal),
Arithmetical Reasoning, Vocabulary, Test of intelligence, etc. They explained various methods of establishing reliability viz. split-half, test-retest, etc. Inter-item correlation, factor analysis for study of validity, methods of calculating correlation coefficient viz. Spearman-brown correlation coefficient, product moment correlation, etc. It also provided the ranges of coefficients of reliability and validity throwing light on expected or usually acceptable ranges of these coefficients. Standardization forms an important part of test-construction and involves ensuring objective administration and scoring of tests, establishing reliability and validity and working out norms. The present study dealt mainly with the objective type tests on Reasoning (verbal and nonverbal), English, Numerical Ability and Clerical Aptitude (speed test) it undoubtedly involved these procedures and therefore the above studies were very much relevant and useful.

In the various research studies reviewed, not a single study was found in which the effect of change in item position on the performance was studied. The present study had considered this additional aspect of the test administration and studied the effect of change in item position.

2.5 Epilogue:

A thorough review of the related literature and research studies was undertaken with a view to obtain information about the methods available and used in different situations and their suitability for the present study. The review of studies provided information about the research work already done and the research work which was not done so far but was necessary to be done.