The process of education includes three major divisions: formulation of objectives, designing education for the achievement of the objectives and assessing the outcomes of education. The function of examinations is to provide such information as is needed for a precise assessment of the outcomes of education. As such examinations are as old as, or rather older than, any system of education. Examinations are of many forms. They may be performance, oral, or written. Written examinations may be classified as objective-type, very-short-answer-type, short-answer-type, and traditional essay-type.

The genesis of traditional essay-type tests may be traced to the Chinese system of administration in vogue around 1115 B.C. The Chinese used the essay-type tests for the selection of candidates for Government offices. Each candidate was shut up in his narrow cell for days and nights and was required to produce compositions in prose and verse on the themes assigned to him. (A detailed description of the Chinese system of testing is given by DuBois, 1964.) It was probably because of lengthy testing time and vast coverage of test material that, as Ebeling (1959, p. 1502) puts it, "the competitive examination system of ancient China ....... helped it serve its primary purpose of providing men of ability for the service of the state." The Chinese system of testing travelled, though not in that strenuous form, from China to the West. By the 19th century, the essay-type tests in the traditional form were well-established in western countries for the award
The use of the essay-type tests in India came as an inevitable consequence of the introduction of the western system of education. Indian universities were modelled on the lines of the British universities; therefore, they also adopted the examination system in practice in the U.K. In India, the essay-type tests were used on a large scale for the first time in 1857 by the universities at Calcutta, Madras and Bombay for the selection of students, and later for the awarding of university degrees. The wide use of the essay-type tests at the university stage gradually led to the adoption of such tests at the lower stages of education also.

Over the years examinations have attained too great an importance in the West as well as in India. People have been branded for life on the basis of their examination results. Not only a man's academic but also his social status has often been judged from his degrees and diplomas. This has led many to seriously question the accuracy of examinations. One of the earliest studies exploring the efficiency of the essay examinations was made as early as 1888 by Edgeworth (1888) in the U.K. Other studies followed suit, of which those done by Starch and Elliot (1913) in the U.S.A., and Hartog and Rhodes (1935) in the U.K. are very important. Thus in the early half of the 20th century, systematic research on examinations was well started.

Advances in various branches of science helped examination research a great deal. Examination experts no longer remained satisfied with guesses. The scientific research approach came to be adopted for the study of almost every problem in examinations. In the beginning, attempts were made to
minimise subjectivity in marking. Many kinds of objective-type tests were worked out. Various scoring methods were evolved to minimise chance factors in scoring. Different statistical concepts, e.g. reliability, validity, difficulty and discrimination indices were developed. Interpretation of examination results also underwent a change. Methods of scaling were evolved for interpreting marks. Test standardisation and establishment of norms completely revolutionised the examination movement. A new discipline known as "Psychometry" (by the British) or "Psychometrics" (by the Americans) was developed, which came to be regarded, as Harper (1960, p. 1) puts it, "both a science (based on rigorous statistical and research methods) and an art (characterised, at its best, by a high degree of technical skill, creative ingenuity, and insight into human behaviour.)"

In the course of research the same experiments were often repeated at different places. This was probably because of a general feeling that since teaching and learning conditions, which vary from place to place, have much bearing on examinations, research findings of one place may not be universally valid. The following statement of the Examination Committee (1962, p. v) alludes to the above feeling:

Methods of testing should be related to practices of teaching and conditions of study, and should be capable of adjustment and adaptation to particular situations and needs.

The same feeling is voiced by Dave (1966, p. 9)

Successive commissions, committees and study groups on education have pointed out the evils of this examination system. The remedies to all these evils cannot be found in the researches done elsewhere in the world, and only careful studies and investigations can offer solutions to problems vexing the people concerned with education.
DEFINITION OF TERMS USED

The two basic concepts of reliability and validity in examinations are used throughout this study and are understood to mean the following:

Reliability:

This refers to consistency throughout a series of measurements. 
(Cronbach, 1960, p. 126.) In other words, "The reliability of any set of measurements is logically defined as the proportion of their variance that is true variance." (Guilford, 1956, p. 436.)

The problem of reliability of the essay tests may be divided into three categories.

Reader Reliability:

The correlation between the marks of two readers is known as 'the reader reliability' of the examination. 
(Gulliksen, 1950, p. 211.)

Content Reliability:

"The reliability of an essay test corrected for attenuation due to the inaccuracy of reading has been termed the content reliability of the essay test." (Gulliksen, 1950, p. 214.)

Total Reliability:

This is a term coined by Harper (1966 a, p. 12) and defined by him as "the correlation between parallel examinations marked by parallel examiners. That is, it is the correlation between examination 1 marked by examiner A
and examination 2 marked by examiner S. This is the reliability coefficient with which we should be most concerned in comparing the reliabilities of essay and objective examinations, as well as in practical decisions of the degree of trust to be placed in an examination mark."

Validity:

"A test is valid when it measures well what it is supposed to measure." (Michaels and Karnes, 1950, p. 104.) To be valid a test must be reliable. But the reverse is not true. A test may be highly reliable, but it may be invalid for the purpose it is supposed to serve.

The following terms have been used interchangeably in the study.

"reader" and "examiner".

"essay test", "essay examination", "essay-type test" and "essay-type examination".

"objective test", "objectie examination", "objective-type test" and "objective-type examination".

"score reliability", "test reliability" and "total reliability".

"grading reliability", "marking reliability", "examiner reliability" and "reader reliability".
A comprehensive and scholarly review of the studies on examinations done prior to 1958 has been made by Basumallik (1959). Unfortunately, the review is marred by a few points, some minor and some major, which may mislead the lay reader. Without in any way detracting from the over-all value of the review, it seems wise at this point to call attention to these points.

While introducing the concept of reliability (p. 128) Basumallik states, "In a general sense, reliability means the accuracy with which a measuring device measures whatever it is expected to measure." (Italics are the present researcher's.) The definition, as it stands, best fits the definition of validity. The italicised portion should better have been replaced by some such as "whatever it does measure". Many researchers would disagree with Basumallik's statement where he says (p. 149):

It is needless to reiterate that in comparing essay and objective test we should not consider reader reliability for obvious reasons. Comparison should legitimately be made between test reliabilities of the two types of examination. The test reliability, for essay examination, should be the content reliability or the reliability estimated by the methods of test re-test, analysis of variance, etc. It is desirable to compare coefficients of a similar nature.

To many it seems that the exact opposite is the case. On comparability of essay and objective tests, Gulliksen (1950, p. 212) observes, "If two parallel essay examinations are matched just as successfully with respect to content as are two parallel objective examinations, the correlation between the two parallel essay forms will practically always be lower than between the two objective forms, owing to the fact that the unreliability of reading will still further lower the correlation between the two
essay forms." Thus, as Harper (1966) holds, for all practical purposes only total reliability, and not content reliability, of the essay test should be compared with reliability of the objective test. Since reader reliability of the objective tests is 1.00 (i.e., the objective examinations completely eliminate this source of unreliability), ignoring it in the essay examinations presents an unrealistic picture.

Test reliability is a function of test length. This Basumallik has also noted in his review (p. 150). Obviously, a 24-hour-or-so essay test (as the Chinese seemed to have) would be more reliable than a half-hour objective test. Hence, for comparing reliabilities of the essay and objective tests the testing time should also be equated. But while giving comparative figures of the essay and objective test reliabilities (p. 150) Basumallik does not record testing times. Explaining the comparative figures of the essay and objective test reliabilities, he (p. 151) states, "The median reliability coefficients, however, are .85 for essays and .93 for objective tests respectively." The direct comparison of reliability figures implies that testing time for all the tests whose reliabilities are under comparison is the same. But his further discussion indicates that testing time varied for different tests. It may be pointed out that direct comparison of the reliabilities of tests of unequal lengths is technically objectionable and is misleading to laymen.

Basumallik further states (p. 151) that "the difference between the two (.85 and .93) is only .8, in favour of objective tests." (Obviously .8 is a printing mistake for .08.) Similarly, at another place on the same page he states that "These reliabilities range from .80 to .96, with a median of .90 which is only .05 greater than the median essay reliability stated
above (i.e. .85). In explaining differences between reliability figures he ignores a vital point that reliability figures do not represent a linear scale. It is $z$ coefficients whose scale is linear. Thus he should have converted the reliability figures to $z$'s, obtained the difference between the two $z$'s, again converted the obtained difference of the two $z$'s to $r$, if he was interested in finding the difference between the two reliability figures. The present researcher doubts that even this approach, though statistically sound, would be able to give any idea about the relative merits of the two tests. The best and probably most easy to understand approach for such comparison seems the application of the Spearman-Brown Prophecy formula. This formula tells us that a test of reliability of .85 has to be lengthened 2.34 times to reach the reliability of .93. In other words, the difference between the reliability figures of .93 and .85 is not nominal, as Basumallik's inference seems to sound. This point should have been made more clear in his review to avoid any misconception in the minds of non-technical people.

Despite the above-noted minor negative criticisms it may again be emphasised that Basumallik's is an excellent review and the quality of his work may not be looked down upon for some minor lapses here and there.

The following conclusions are drawn from his review:

One must say that reliability of a usual essay-type examination - reader as well as content - is not as high as that of an objective test. This is true even of subjects like mathematics and chemistry. An objective test, even when imperfectly constructed, is not likely to have reliability lower than .70, whereas the usual essay test will have still lower reliability. It may be a problem for further experimentation to find out how best an essay
test can be improved. Under certain conditions, e.g., team impressionistic marking, however, an essay examination can be so improved as to stand fair comparison with an objective test in point of test reliability. But these conditions are not always feasible to employ in actual situations involving a huge number of examinees and requiring quick decision. More often we have to depend upon a single marker in a real situation, and the marking of a single examiner is found to be considerably erroneous. On the whole, the research on essay reliability, done so far, presents incomplete and controversial evidence on many issues. Intensive research is needed to see how far it is possible to improve the reliability of an essay.

Basumallik has reviewed the earlier studies so well that it did not seem necessary to repeat this work. The present researcher, therefore, confined himself to reviewing later studies as well as some important studies omitted by Basumallik.

EXAMINATION RESEARCH ABROAD

1. Conversion of Essay Items to Objective Items

Some foreign and Indian experts (Vernon, 1940; Wrightstone, Justman and Robbins, 1956; Thorndike and Hagen, 1961; Examination Committee, U.G.C., India, 1962; Standards Committee, 1965) hold that the essay-type tests measure those larger outcomes of education which may not be measured by other paper-and-pencil-type tests. Others (Sims, 1948; Stalnaker, 1951; Noll, 1957) hold that this claim of the essay-type tests is generally accepted without proof, evidence, or supporting logic.
Implicit in the claim of essay-type tests is that objective tests do not and cannot measure the larger outcomes of education. Some experts (Ballard, 1923; Diederich, 1957; Leo Nedelsky, 1957; Engelhart, 1957; Ryans, 1958; Frederiksen, 1960; Kelly, 1963; Harper, 1963 (b); Hubbard, 1963; Hill, 1964) hold that most of the educational outcomes measured by the essay tests can be measured by good objective tests. Ebel (1953) and Anastasi (1966) hold that the impression that the objective tests can measure only trivial educational objectives is probably due to the fact that such objective tests can easily be prepared by un-skilled writers.

However, some studies have tried to find out empirically as to what extent the material covered by the essay tests can be made objective.

Sims (quoted in Woods, 1953) showed that out of 243 essay questions in elementary education examinations, 103 were simple recall, 97 short answer and only 43 discussion-type questions. In secondary education examinations, out of 215 questions, 54 were simple recall, 63 short answer and 98 discussion-type questions. He concluded from his study that only "30.5 per cent of the questions (were) in reality .. subjective questions and better than two thirds of the grand total studied could have been converted to objective-type tests".

Orleans and Sealey (quoted in Woods, 1953) while not making a statistical study of the problem, did call attention to it by giving an essay test and then converting it to an objective-type test to show the thesis of objectivity in essay questions.

Woods (1953) found that in one hundred separate essay tests studied, there were a total of 1,456 answer items expected as shown by the teachers
suggested answers. Of that number, 1,422 were capable of conversion to the objective-type tests. The study suggests that if the essay-type questions which can be converted to the objective-type tests are handled thus, there would be much saving of teacher-pupil time, and students would be better able to understand the fairness of the grading. The study also suggests that teachers use the essay tests where objective tests would have done the job better, and teachers frequently do not accomplish the expressed goal of the essay tests of measuring pupil's thinking, organisation and expression due to improperly worded questions.

Moore (1954), giving a concrete example, has shown how one essay question (for graduate medicine) can be covered by seventeen objective items. Assuming that this example is average, he concluded that a five-question essay examination may be covered in eighty-five objective items. The usual objective examination of two-to three-hour duration contains two-to three-hundred items, which could correspond to twelve to eighteen essay questions. Thus, the objective examinations can cover a broader spectrum of knowledge than the essay examination in the same time period.

2. Reliability and Validity of the Essay Tests

Several investigators have explored the reliability and validity of essay-type tests.

Wiseman (1956) had 30-minute compositions written by a group of 173 students and graded by four examiners. This exercise was repeated with another title four months later, and the essays were graded by the same four examiners. Correlation between the aggregate marks on the first and second essay was .89;
corrected for the range of ability in the sample, it was calculated to be .92. This means that the total reliability of the essay tests is high for the pooled marks of four readers.

Wiseman, in a validity study on a different sample (N=117), correlated entry test scores including an essay with the following five criteria:

1. Total school certificate results.
2. School Certificate English Language.
4. Teachers' over-all estimate.
5. Teachers' estimate of written English.

Multiple r's calculated for the whole battery, and for the battery without the essay mark included, revealed that the addition of the English essay increased the size of the coefficient in the case of all the five criteria. However, it may be added that the design of this experiment is open to criticism on the following grounds:

Selection of the sample was biased in favour of students regular in their studies. It is also probable that school certificate standards differed significantly from school to school, because all the schools selected for the study were not of the same quality. Hence the inference based on the correlations with school certificates may be misleading. Thus, the findings of this study may only be considered tentative pending corroboration by further evidence.

Penfold (1956) explored variability in marking essay tests. Five-minute essays written by 16,000 candidates were marked by impression by sixteen
examiners, each examiner marking one thousand scripts. After an interval of time (the exact period is not reported) 165 scripts selected at random from the whole batch were all re-marked by fifteen of the same examiners, the sixteenth not being available. The results indicated that "there was significantly high variation between different examiners' standards of marking and also between the marks of the same examiner on the two different occasions of marking."

From the above results Penfold anticipated that in all probability the best chance of improving consistency would be to devise an analytic system of marking longer essays. He, therefore, studied the analytic marking of 20-minute essays. Sixteen examiners were involved in marking, including the chief examiner. At a standardising meeting, the examiners discussed the details of marking. To be quite sure of the conformity in marking, each examiner marked a photostat set of 25 scripts judged by the chief examiner to cover the whole mark range. Finally, after a prolonged period of marking to this scheme, the extent to which the errors had been reduced was checked by an analysis of marking of twenty more photostatted essays. The results suggested significantly high variations in marking between the examiners, in spite of the extensive training and experience.

It may be observed that no examining body can afford to give that much intensive training to its examiners as was given in this experiment. This is another evidence to prove that variations between the markings of examiners is an inherent and incorrigible weakness of the essay-type tests.

Pidgeon and Yates (1957) found test re-test reliability for the essay tests to be .77 if the same examiner re-marked, and .72 if a different
examiner re-marked. In a follow-up study of 473 students, who had completed their secondary school course, they found that the old-type (i.e., essay-type) English examination provides significantly less satisfactory forecast of subsequent success in secondary schools than is obtained from the results of an objective test.

Hudson (1960) found that (1) neither the Cambridge F.R.S.'s nor the Cambridge D.Sc.'s had better degree classes than those of their respective control groups; and that (2) although Cambridge F.R.S.'s were roughly three times as numerous as those from Oxford, they were far less likely to have a first class degree. He concluded that "if our result is accepted as approximately accurate, it would appear that class in Tripos bears no direct relation to subsequent eminence in scientific research as assessed by either of our two criteria."

Valin (1961) found total reliability of the essay tests, by correlating marks on two tests, to be .12 for Arabic Philosophy, and .52 for Mathematics. Valin has not reported testing time. As such, it is not clear whether the very low reliability of Arabic tests was due to short lengths of the tests or due to subjectivity in the essay type tests.

French (1961) got 300 short essays written as homework by college freshmen, marked independently by ten English teachers, nine social scientists, eight natural scientists, ten writers or editors, nine lawyers and seven business executives, on nominal compensation. The examiners were asked to use their own judgement in marking, and sort the papers into nine piles in order of merit. The average correlation between all the examiners was .31. Out of the 300 essays, 101 received all nine different grades, and no paper
received less than five different grades. The average correlation between English teachers was .41. The higher correlation among English teachers was attributed to their less erroneous markings. He concluded that nothing at all is being measured very well by the essays.

Many teachers feel that reader reliability may be improved if essays are marked on generally accepted criteria. Postvedt (1965) explored the criteria for evaluating high school English Composition. Using generally accepted criteria, thirty experts gave numerical ratings to twenty themes. He concluded that perhaps the only conclusion justified by the study was that, although teachers of English composition might feel that criteria are important in evaluating theses, there was no evidence of consistency in the employment of such criteria.

Lehtovara (1966) in a study on a sample of 1330 students concluded that the validity of the essay examination could be improved by the modern test theoretic methods. The reliability of the grading of test papers was found to be .85. This is one of a very few western studies where reader reliability of essay examinations was found to be above .8. It may be recalled that the grading reliability of .85, which looks very high, is far below the reader reliability for the objective-type tests. For an objective test scored carefully with an agreed key, the reader reliability is 1.00. Even for careless scoring, the reader reliability of the objective tests is far higher than .85. Ebel (1965 p. 312) remarks: "Even if over half of the answer sheets in a set of thirty-five were scored wrongly, with errors like those shown below, the coefficient of scoring reliability would still be closer to .99 than to .98.
<table>
<thead>
<tr>
<th>Size of Scoring Error</th>
<th>Number of Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>35</td>
</tr>
</tbody>
</table>

It may be stated that a test of reliability of .85 has to be lengthened approximately nine times to reach the reliability of .98.

Myers (1966) had essays of over 80,000 students double marked. In a period of five days, 145 readers gave each of the essays two independent readings on a four point scale. The average reliability of the judgements calculated by analysis of variance method among all readers and across all papers ranged from .26 to .49 over a five-day period.

One of the arguments constantly brought up in favour of the use of the essay examinations is that they purport to measure creativity. However, Page (1967) in a study of a sample of essays written by secondary school students judged by independent examiners found the measurement of creativity on essay tests to be the least reliable attribute rated by human judgement, and mechanics the most reliable.

Coffman (1966) investigated the validity of the essay-type tests to predict writing ability. The students who were the subjects of the research were in classes XI and XII in 24 American secondary schools. Each student
took six objective and two semi-objective tests and wrote five essays of different lengths. Each student wrote on five different topics and each essay was marked by five trained examiners. The sum of these twenty-five marks was the student's criterion score. The reader reliability of one essay read once, that is, the correlation between two independent readings of one essay ranged from .11 to .34, the average being .26 which is a very low figure. The validity coefficient of essay tests ranged from .33 to .55. He concluded that "in order to obtain validity coefficients comparable to those obtained from an one-hour objective test using only essay tests, it would be necessary to assign at least two topics to each student and each read by five different readers, or to assign three topics and have each read by three different readers." Vernon (1940) makes exactly the same recommendation for attaining a high reliability for the essay-type test; that is, to have a minimum of three essays from each pupil, each marked by three markers. A further analysis of the data of Coffman et al. (quoted in Hill, 1967) showed the following major findings:

1. "The reliability of essay scores is primarily a function of the number of different essays and the number of different readings included." These factors are more important than the length of the essays, and they have more value than analytical reading.

2. "Validity is lost when optional topics are given." In a situation where choice is allowed, "a student's rating might depend more on which topic he chose than how well he wrote."

3. "When objective questions specifically designed to measure writing skills are evaluated against a reliable criterion of writing skills, they prove to be highly valid."
4. "The most efficient predictor of a reliable direct measure of writing ability is one which includes essay questions in combination with objective questions. A combination of two objective tests and one essay, requiring a total testing time of one hour, correlated most highly with the criterion. However, the validity added by a twenty-minute essay is small, and it requires that the essay be marked by at least three examiners. It is doubtful that the slight increase in validity alone can justify the increased cost."

Bracht and Hopkins (1968) found reliability coefficients of two 20-minute objective tests, calculated by Formula 20, to be .40 and .50. The reliability coefficients of three essay tests of the same time length, calculated by the analysis of variance method (that is content reliability), was found to be .58, .26 and .69. As already observed, the content reliability of the objective test is comparable to the total reliability of the essay-type test. If reader reliability were taken into account, the essay test reliabilities reported would be reduced considerably, while the objective test reliabilities would remain unchanged. Thus the essay and objective test reliabilities as reported here should not be considered as directly comparable. The other major findings of this study are as follows:

Cumulative grade point average correlated very significantly with performance on both types of test, but the correlation with the objective tests was significantly higher. Writing ability was significantly related to both the essay and objective tests, the difference between the correlations being not significant. The findings of this study do not support the common supposition that the objective and essay tests measure different abilities. In addition, there is clear evidence that the objective test
Klein and Hart (1968) investigated the chance and systematic factors affecting essay grades by getting the scripts of over 1500 students of Law marked by seventeen professors. They found that students tended to be consistent in their performance in different essay questions used in law school. The agreement in grading among the professors was .76.

3. Conclusions Suggested by the Foreign Studies.

The following conclusions may be drawn from the studies done abroad:

1. Most of the material covered by essay tests may be converted to objective tests. Much more content area may be covered in testing if essay tests are converted to objective tests.

2. For short essays, reliability is very low.

3. There is significantly high variation between essay markings of different examiners, or of the same examiner on two different occasions. Intensive training for standardising markings seemed to have little effect in reducing the variations in markings of different examiners.

4. Creativity, which is supposed to be the major quality assessed by the essay test, is least reliably graded.

5. In essay examinations in order to get a validity coefficient comparable with that of a one-hour objective test, it is necessary that at least two topics each marked by five independent examiners or three topics each marked by three independent examiners be obtained.
6. The essay and objective tests do not necessarily measure different abilities. On the other hand, the evidence shows that objective tests are significantly better measures of academic achievement.

7. Validity is lost when optional questions are provided, and a student's score may depend more on what topic he wrote than on how well he wrote.

8. The predictive validity of essay tests in forecasting success in succeeding examinations is significantly less than that of objective tests.

9. Essay examination results are not valid measures for predicting eminence in scientific research.

EXAMINATION RESEARCH IN INDIA.

Many areas in examinations have been covered by research studies in India. These studies can be broadly classified in two sub-heads: Functional aspects of examinations and efficiency aspects of examinations.

1. Functional Aspects

Studies on the functional aspects of examinations have produced what is essentially a survey of the statistics about different examinations in India. A Mysore University survey (1965) showed that pass percentages for Pre-University examinations vary from university to university in India. In 1965 the pass percentage for the Pre-University examination for the Mysore University was 30, whereas it was 55 for the Bombay University. The pass percentage for the same examination even for the same university varies from
year to year. The pass percentage for the Intermediate examination varied from 26 in 1935 to 48 in 1945 in the Mysore University. Mukerjee (1954) in a study done in the West Bengal found that the pass percentage for the Matriculation examination varied little from 1912 to 1945. He inferred that pass percentages remain almost static from year to year. Dave and Patel (1966), DEPSE (1962 a, and undated a,b ) and the S. S. C. Board, Maharashtra (1963 a) found that the pass percentages of regular students are markedly higher than those of the private candidates. The S.S.C. Board, Maharashtra (1960) found that there was better pass performance among urban students than among their rural colleagues. They found that at the age of 19-20 years girl students do better than boy students, while at the age of 16-19 years boy students do better than girl students. It was also noticed that the mean percentage of successful candidates rapidly declines with the advance of age above 16 years. This tendency was found both for boys and girls. The regional languages were found to be more efficient as the medium of instruction than the English language in enabling the students to score higher marks in Science, History and Geography.

Chitrakara (1961), DEPSE (undated, a), Deshmukh (undated), Bhanot (1961), Deshmukh and Kamat (undated), The M.S. University, Baroda, (1964), the Mysore University (1965), Pillai (1965), and the S.S.C. Board, Maharashtra (1960, 1963) studied the wastage and stagnation among students. The wastage and stagnation were found for each class to be between 40% to 70%. The general causes of the wastage and stagnation as inferred by these studies are poverty, lack of proper atmosphere at home and school, want of equipment and qualified staff in schools. Lele et al. (1966) analysed the failure rates of the Preparatory Commerce examination of the M.S.University, Baroda for the years from 1950-51 to 1964-65. (The Preparatory Commerce examination is a
one-year course where students are admitted after passing the S.S.C. Examination. They found that there was a significant upward trend in the failure rate in the Preparatory Commerce Examination. Lele et al. inferred that assuming that the quality of the S.S.C. examination, as evinced by the S.S.C. examination marks, remains the same, the Preparatory Commerce examinations have become more difficult.

Mitra (1956) and Misra (1968 a) factor analysed examination marks of students to find out common factors underlying different academic subjects. Mitra's study does not provide any conclusive evidence, probably because the number of cases in the sample was very small. Misra (1968 a) found three factors in the Matriculation examination of the Gauhati University. The first factor, which was dominant, was named "verbal". The second and third factors were almost of equal order, and were named "problem-solving" and "memorisation". Contrary to his expectation, Misra did not find a numerical factor. He stated as the probable cause that the condition of having at least two variates for each ability was not satisfied in his experiment in the case of the numerical factor. Lele et al. (1964) factor analysed the S.S.C. and P.Sc. examination marks, the two being successive examinations, for the years 1957-58, 1958-59 and 1959-60. They found two common factors, namely, "verbal" and "numerical".

Ray (1969) found that the failure rate in Mathematics in H.S.L.C. examination is on increase. It has shot up from 45.7% in 1966 to 76.9% in 1968. He attributes the reason of failure to the poor quality of students appearing in the H.S.L.C. Mathematics examination.

The studies on the functional aspects of examinations provide useful information, but their number is very few. On some points the findings are
contradictory, which suggest the need of further investigation. Moreover, the value of such studies is limited as regards the assessment of the effectiveness of examinations as a tool of measurement.

2. Efficiency Aspects.

To be an efficient tool of measurement, an examination should be valid, precise and economical. Studies on the efficiency aspects of examinations may be classified in the following sub-heads: variation in marking, marking errors, reliability of marks and validity of marks.

1. Variation in Marking

(a) Variation for the same set of scripts

There are several studies on the variability of marking the same set of scripts by different examiners, or by the same examiners on different occasions.

Mukerjee (1961) had two essays, one in Assamese and the other in English each marked independently by five competent examiners. The five examiners for the Assamese scripts awarded marks ranging from 14 to 29. The five examiners for the English scripts awarded marks ranging from 6 to 22. The two scripts were marked by the same examiners after a week. The highest difference in marks given by the same examiner to the same scripts on the two occasions was 12 (i.e. 31-19) for Assamese, and 11 (i.e. 22-11) for English. The mean marks awarded to the Assamese scripts varied from 26.4 to 18.6, and for the English script from 15.6 to 8.4 on the two occasions. As is evident from the study, not only did the examiners differ from each other; they even
differed from themselves on the two occasions-- i.e. after a lapse of seven days.

George (1964) studied the difference in examiners' markings in the Pre-University Test of the Kerala University by selecting a random sample of 40 scripts each of English I, English II, Mathematics, Physics, Chemistry, Botany and Zoology. These scripts were marked independently by two separate external examiners of the University. Except for English I, and Zoology, in all other subjects the differences between the markings of the external examiners were found significant at the 1% level. In Zoology the difference was found significant at the 5% level.

George (1964) in another experiment had a large number of scripts marked by two independent University examiners in the above-noted subjects of the P.U. students of the Kerala University. He found that the percentage of disagreement in classifying students in different categories, that is, I, II, III and Fail, ranged from 22.6 in Mathematics to 52.2 in Botany. In Physics one student who was failed by one examiner was given a first class by the other examiner. In Botany there were two such students. In Mathematics one student who was given a fail and twelve students who were given a third class by one examiner were given a first class by the other examiner.

It may be observed that the scripts used in both the experiments of George (1964) seem to have contained crosses and ticks given by the original examiners. The crosses and ticks might have invited the attention of the other examiners to the points marked by the original examiner. Thus, the obtained disagreement in the markings should be considered as the minimum obtainable in a real examination situation.
Taylor, Tluanga and Misra (1966 b) had answer scripts of 100 examinees of B.A. Part I English examination of the Gauhati University cyclostyled and examined independently by nineteen examiners. The mean marks ranged from 19.2 to 35.0 (out of 100 marks) and pass percentages ranged from 7 to 62 for the nineteen examiners. In some cases the disagreement in awarding marks was extreme. For example, it was found that a student who got a 3rd rank from one examiner was given a 98th rank by another. These scripts were originally marked by the University examiners. Out of the 100 students, the number of original failures who passed and the original passes who failed in the second marking ranged from 20 to 37 for the nineteen examiners. It may be observed that in the experiment the copies of scripts were not reproduced in the students' own handwritings. They were typed. Thus, disagreement in marking due to the quality of handwriting was eliminated. As such, the disagreement as reported in the study should be considered as the minimum expected in a real situation. Harper (1967 a) had facsimile copies (by Xerox process) of 10 scripts of History examination of class X of a Board of Secondary Education marked by 90 experienced examiners of the same Board. It was found that the only student considered fit for the award of a star (i.e. above 75% marks) was also failed by several examiners. The script which was given first or second rank by almost all the examiners got from a high first class to a bare pass mark. The variation in pass percentages ranged from 10 to 80. This is the only study in India where variation in markings was explored in an almost real examination situation, i.e. the scripts were reproduced in students' own handwriting, these were sent to examiners during the time when they usually examine the Boards scripts, and the examiners were paid remuneration at the Board's rate. Thus, the findings of this study may be considered as representative of variability in markings.
Harper et al. (1967) in another study had scripts in Hindi, History, Geometry and Biology of Class X final examination of a Secondary Board re-marked, some by the same examiner and some by another examiner. The highest variation in marks was found in Geometry (supposedly an "objective" paper). The same examiner, re-marking his own scripts differed from himself by 42% marks. The highest mean difference was found in Biology (another supposedly "objective" paper), a mean difference of 9.8 for ten examiners each re-marking 50 of their own scripts. They found that the variability in marking remains almost of the same magnitude whether an examiner re-examines the scripts previously marked by himself or marked by another examiner.

Harper et al. (1967) analysed the variability in marking in a different way also. In the above-noted experiment the markings of the two examiners were equally erroneous. They, therefore, selected one examiner who was considered by the Board to be highly experienced in marking, that is, a "standard" examiner. The scripts marked by the "standard" examiner were re-marked by other experienced examiners of the Board. The mean variations in the two markings were found to range between 14% to 32% of marks in History, 12% to 24% in Hindi, 12% to 16% in Biology, and 16% to 28% in Geometry.

Many teacher feel that marking is objective in Science and Mathematics papers. The experiments of George (1964) and Harper et al. (1967) show that in essay-type tests subjectivity of markings in the Science and Mathematics papers is no less, if not more, than that in the Arts papers.

Taylor (1962 e) had scripts of 45 students in English (two papers) Economics, History, Logic and Mathematics independently marked by two
examiners. The highest difference between means was found in logic, where
the mean mark of one examiner was 55.8, and of the other examiner was 46.0.
The lowest difference between means was found in History, where the mean
marks varied from 40.4 to 39.5. The highest difference in the standard
deviations was in Logic, where the s.d.'s varied from 18.0 to 10.8. He
concluded that "the experiment makes it clear that an examination mark has
neither the sanctity nor the precision which is usually attached to it". He
recommended the use of scaling to minimise the effects of errors in
marking. Some people suggest that to improve the reliability in marking essay-type tests, fewer number of scripts should be given to examiners. This experiment makes it clear that even for a very few scripts the differences in means and s.d.'s in the markings of two examiners are alarmingly large.

(b). Variations for equivalent sets of scripts

Taylor did several experiments on the variation in marking for statistically equivalent sets of scripts. He observed that, "by allotting roll numbers at random, or otherwise mixing the candidates one can ensure that the sub-groups are statistically equivalent". He used this method of making sets of scripts statistically equivalent in the following three studies, so that any differences in the mark distributions can be attributed to the examiners and not to the candidates.

Taylor (1962 b) found that in an English paper of college students, one examiner produced a distribution which Taylor describes as S1 (5, 3)
(where S1 is the mean, 5 the difference of the upper quartile from the mean, and 3 the difference of the lower quartile from the mean), and the other one
for the equivalent set of scripts produced the distribution 31 (3, 4). With one exception, all the marks given by the second examiner were lower than any marks given by the first examiner.

Taylor (1962 b) had a first year college paper in Biology, involving 210 candidates, marked by two pairs of examiners, A and B taking 110 scripts, C and D taking the remaining 100 scripts. The two sets of scripts had been made statistically equivalent by allotting random roll numbers. The marks of each pair of examiners were averaged. The median mark was 66.5 for A and B, and 42.0 for C and D. The true means were respectively 64.9 and 40.4.

Taylor concludes:

The difference in the average marks (whether one uses the median or the mean) is thus 24.5. With A and B, 88 per cent of the candidates got 55 marks or more; with C and D on the other hand 95 per cent of the candidates got less than 55 marks. The difference is even more remarkable in that each total mark is found by combining the separate marks of two examiners, a procedure which would tend in general to smooth out differences in the standards of marking. The examiners were unwilling to admit even the possibility of such large differences until the analysis was put before them.

Taylor (1963 a) found the pass percentages of the equivalent scripts marked by different examiners varying from 11% to 96% in English, 27% to 77% in Mathematics, and 15% to 70% in Science, in the Matriculation examination of the Gauhati University.

II. Misclassification of Students due to Marking Errors

Some people feel that classification in divisions is reliable though the marks may not be so. Some Indian studies have explored this point.
Taylor (1962 c) studied a hypothetical group of one thousand students with an average mark of 45 and a standard deviation of 15. He assumed that a pass was secured by a mark of 30, second class by a mark of 48, and a first class by a mark of 60. He further assumed the standard error of marking of 5 marks. He found that about 25% of the students would be wrongly classified in I, II, III and Fail categories in the above-noted case. He also found in this case that 9% of the students who really deserve a pass would fail on a paper by accident. On this assumption, he calculated the probability of passing a student in a ten-paper examination when the student has the same ability in all the papers and would pass the whole examination on the basis of his true mark. The probability of passing simultaneously in a ten-paper examinations was (for such a student) \(0.39\).

Harper (1963 b) showed that due to the provision of choice, students of the same ability may get different marks. Taking an example where two students of the same ability appear for a question paper in which five out of nine questions are to be answered, he demonstrated how one student may get 75 marks (a distinction) and the other may get 65 marks (a first class) out of 100.

Harper et al. (1967) got 1,000 scripts of class X examination in each of Hindi, History, Geometry and Biology double marked. They found the expectancy of getting the same division if marked again as follows: In History 9% for I division, 42% for II division, 64% for III division, and 63% for fail. In Hindi 36% for I division, 53% for II division, 66% for III division, and 67% for fail. In Biology 39% for I division, 62% for II division, 64% for III division and 63% for fail. In geometry 64% for distinction, 66% for I division, 65% for II division, 61% for III division,
and 90% for fail.

George (1964) got 3,416 scripts of English, Mathematics, Physics, Chemistry, Botany, and Zoology of P.U. students of the Kerala University remarked. He found that "results of nearly 30% of the students would have been altered in terms of pass divisions or failure either to their advantage or disadvantage by re-valuation." He came to the conclusion that "the present system of evaluation produces considerable degree of error".

III. Reliability

Studies on reliability have been grouped under three heads: Content Reliability, Reader Reliability and Total Reliability.

(a) Content reliability

As defined earlier, the reliability of an essay-type test corrected for attenuation due to the inaccuracy of reading, has been termed the content reliability of the essay test.

There are studies by Gayen et al. on the content reliabilities of certain papers of class X examination of the Board of Secondary Education, West Bengal. Gayen et al. do not make any differentiation between various types of essay test reliability. They have used only one term "reliability" for the various types of reliability reported by them. The studies reported under the head "Content Reliability" have been grouped by the present researcher, as they seemed to him to fit the conditions of the content reliability.
Gayen et al. studied the content reliabilities for several papers of class X examination of the Board of Secondary Education, West Bengal.

In the first study (1961) the content reliability was found to be .63 for Mathematics. In the Second study (1962) it was found to be .7 for English I and II each, .6 for English III, and .85 for all the three papers taken together. In the third study (1966) it was found to be .85 for Sanskrit, .80 for Hindi I, .68 for Hindi II, .64 for Hindi I and II taken together, .74 for Bengali I, .85 for Bengali II, and .70 for Bengali I and II taken together. In the fourth study (1967) it was found to be .8 for Physics I and II each, .83 for Physics I and II taken together, .75 for Chemistry I and II each, and .86 for Chemistry I and II taken together. In the above studies each paper, except English II which was of two hours, was of three hours.

It may be observed that in India the above-noted studies done by Gayen et al. are the pioneering studies analysing a Board's results by modern psychometric methods. The strengths of these studies are many: There is no study in India on Public examinations where the sample is so representative of the population as in these studies. These are the only studies in India which have tried to analyse every aspect of examinations using a wide variety of statistical tools of analysis. Of course, from a pioneering study, one cannot expect to get answer to all and every problems in the field. It is also natural that even such elaborate studies may have some weak points. Without in any way detracting from over-all
value of the studies, it appears desirable to invite attention to the following points:

Their definition of difficulty index is novel i.e. different from generally accepted definition of difficulty index. Unless one knows the special meaning given to this term by them — and it is very unlikely that a technical man would consult their definition due to his confidence in the knowledge of the term — he will certainly draw wrong conclusions from their figures of difficulty index. Commenting on the generally accepted definition of difficulty index, Ebel (1965, p. 359) rightly states that "This inverse relation between what is ordinarily meant by difficulty and the index used to measure it is illogical, but the convention has become so firmly established that to attempt a change would be only to contribute to confusion."

While discussing the concept of difficulty index, Gayen et al. stated that the difficulty of an item depends upon the ability of the group who answered the item. But in the calculation of the difficulty index of an item, they seem to have ignored the ability level of the students answering the particular item.

In splitting the Mathematics paper in two halves they almost ignored the matching of course content, and were mainly guided by the consideration of having the other half which could best correspond with the marks awarded in the first half. This approach may lead to spurious reliability. In the application of \( \text{KIB}_{20} \), they have treated the total of marks awarded on the several sub-groups of a question as the item score for that question. This approach would also lead to spurious reliability. However, they have amply made it clear in a true researcher's spirit that these were the only approaches
which they thought they could possibly adopt, though they knew that their data did not satisfy all the conditions needed for such approaches.

Despite these minor criticisms, it may be emphasised that Gayen et al. have set the tone for analysis and interpretation of many problems of essay-type examinations, and their endeavour deserves all commendation. To quote Shakespeare:

Frail creatures are we all, To be the best
Is but the fewest faults to have

Misra (1968 a) investigated content reliabilities of English I, II and III, History, Geography and Mathematics of the Matriculation examination, 1963 of the Gauhati University by factor analysing the marks obtained by 1,241 students selected from the examination on the basis of a random sampling. He found the content reliability (obtained $h^2$) to be .76 for English I, .63 for English II, .63 for English III, .60 for History, .49 for Geography, and .50 for Mathematics.

Misra (1968 b) in another study found the content reliability of English I of three-hour duration of the Pre-University examination 1966 of the Gauhati University by analysis of variance, and split-half methods to be around .8.

Harper et al. (1967) investigated the content reliability of Hindi, History, Biology and Mathematics of class X examination of a secondary board of education. They found the content reliabilities for various examiners in Hindi ranging from .45 to .77; for History ranging from .21 to .76; for
Biology ranging from .41 to .64; and for Mathematics from .68 to .90. They attribute the variability in the content reliability of the same question paper for various examiners due to the fact that different examiners may emphasise different aspects of the answer, e.g. style, factual accuracy, organisation, creative thinking, etc. Since the proportion of each of these in the content of answer scripts may differ, the emphasis on different aspects is likely to produce different content reliabilities for different examiners.

It may be noted that Gayen et al. (1961, 1962 and 1966), Misra (1968 a and 1968 b) and Harper et al. (1967) found approximately the same reliability figures for different papers, except Mathematics. For the Mathematics paper Harper et al. (1967) report high reliabilities, i.e. .68 to .90 (median .77), and Gayen et al. (1961) and Misra (1968 a) report low reliabilities, i.e. .63 and .50 respectively. One of the probable reasons for such difference in the reliability figures of Mathematics may be that Harper et al. (1967) calculated reliability for Geometry only, whereas Gayen et al. (1961) and Misra (1968 a) calculated reliability for the paper containing questions on Arithmetic, Algebra and Geometry. It is possible that Arithmetic and Algebra tests may be less reliable than a Geometry test. Thus, the inclusion of Algebra and Arithmetic in the question papers studied by Gayen et al. (1961) and Misra (1968 a) might have resulted in the low reliabilities obtained by them. However, this point needs further exploration.

(b) Reader reliability

As defined earlier, the correlation between the marks of two readers (examiners) is known as the reader reliability of an examination.
Taylor, Tluanga and Misra (1966 b) found the reader reliability of English I of B.A. Part I (of three hours' duration) of the Gauhati University by getting a random selection of 100 scripts (cyclostyled copies) independently marked by nineteen examiners. The average correlation among the examiners was found to be .77 and the coefficient of concordance ($\tau$) was found around .75. These scripts were previously marked in the University examination. The coefficient of correlation between the University marks and the average of the nineteen examiners' marks was found to .67. The low correlation of the University marks with the average of the nineteen examiners' marks seems to suggest that marking in real examinations is more erroneous than in experimental situation.

Harper (1967) found reader reliability of a History paper of class I of a Secondary Board of Education by getting facsimile copies (by Xerox process) of ten scripts, fairly representative of all levels of ability, marked by ninety experienced examiners of the same Board. The average reader reliability was found of the order of .83.

Harper et al. (1967) found the following figures of reader reliability for different subjects of class X examination of a Board of Secondary Education.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Range of reliability figures</th>
<th>Median of reliability figures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>.61 - .94</td>
<td>.79</td>
</tr>
<tr>
<td>History</td>
<td>.49 - .93</td>
<td>.77</td>
</tr>
<tr>
<td>Biology</td>
<td>.67 - .91</td>
<td>.80</td>
</tr>
<tr>
<td>Mathematics</td>
<td>.81 - .99</td>
<td>.96</td>
</tr>
</tbody>
</table>
It may be observed here that reader reliability figures reported in Indian studies compare well with such figures reported in foreign countries. For example, as reported earlier, reader reliability was found to be .31 by French (1961) and .85 by Lehtovaara (1966). This shows that Indian examiners are not less efficient than their counterparts in most advanced countries.

(c) Total reliability

As defined earlier, the total reliability is the correlation between the marks given by two parallel examiners on two parallel essay tests. The present researcher is not aware of any study in India which strictly satisfies the conditions for the determination of the total reliability of the essay test. However, there are studies comparing the marks obtained by a group of students on internal and external examinations, both of which are supposed to be measuring the same content. These results may be roughly taken as an estimate of the total reliability of the essay-type tests. It may be noted that the studies reported under this head have not used the term "total reliability". They have been grouped by the present researcher under this head, as to him they seemed to satisfy the conditions of the total reliability.

Gayen et al. (1962) found the following correlations between the school test examination marks and Secondary Board examination marks of the same group of students in different school subjects.

<table>
<thead>
<tr>
<th>Subject</th>
<th>N</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>English I</td>
<td>507</td>
<td>.61</td>
</tr>
<tr>
<td>English II</td>
<td>636</td>
<td>.63</td>
</tr>
<tr>
<td>English III</td>
<td>634</td>
<td>.44</td>
</tr>
<tr>
<td>English Total</td>
<td>494</td>
<td>.73</td>
</tr>
</tbody>
</table>
The S.S.C. Board, Maharastra (1960) did a similar study by correlating marks of school test examination with class X final examination marks. For the purpose of sampling, the examination area was divided in two regions. A fairly random sample was drawn from each region. The results obtained are reported below:

<table>
<thead>
<tr>
<th>Subject</th>
<th>I Sample</th>
<th>II Sample</th>
<th>N</th>
<th>Reliability</th>
<th>N</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>382</td>
<td>566</td>
<td>.49</td>
<td>.51</td>
<td>.47</td>
<td></td>
</tr>
<tr>
<td>Higher Level</td>
<td>397</td>
<td>562</td>
<td>.57</td>
<td>.47</td>
<td>.52</td>
<td></td>
</tr>
<tr>
<td>Social Studies</td>
<td>255</td>
<td>408</td>
<td>.46</td>
<td>.52</td>
<td>.54</td>
<td></td>
</tr>
<tr>
<td>General Science</td>
<td>374</td>
<td>500</td>
<td>.54</td>
<td>.62</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Mathematics</td>
<td>155</td>
<td>310</td>
<td>.79</td>
<td>.68</td>
<td>.68</td>
<td></td>
</tr>
<tr>
<td>English (with test)</td>
<td>327</td>
<td>432</td>
<td>.69</td>
<td>.69</td>
<td>.68</td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>237</td>
<td>263</td>
<td>.57</td>
<td>.60</td>
<td>.68</td>
<td></td>
</tr>
<tr>
<td>Classical Language</td>
<td>224</td>
<td>332</td>
<td>.69</td>
<td>.68</td>
<td>.68</td>
<td></td>
</tr>
</tbody>
</table>
Harper (1966 a) has statistically shown that the total reliability of an essay-type test is equal to the product of its reader and content reliabilities. Accordingly Harper et al. (1967) multiplied the figures of the content reliability calculated by analysis of variance method with the figure of the reader reliability obtained for the same set of scripts. The median total reliabilities for various subjects were found as noted below:

<table>
<thead>
<tr>
<th>Subject</th>
<th>Total Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>.58</td>
</tr>
<tr>
<td>History</td>
<td>.35</td>
</tr>
<tr>
<td>Biology</td>
<td>.50</td>
</tr>
<tr>
<td>Mathematics</td>
<td>.72</td>
</tr>
</tbody>
</table>

Following Harper's approach, that is, by multiplying the figures of the reader and content reliabilities, Misra (1968 b) found the total reliability of English I of P.U. Examination (of three-hour duration) of the Gauhati University to be .64.
# Table showing reliability figures obtained in Indian studies.

<table>
<thead>
<tr>
<th>Content Reliability</th>
<th>Maths</th>
<th>Eng. I</th>
<th>Eng. II</th>
<th>English III</th>
<th>English I+II+III</th>
<th>Sanskrit</th>
<th>Hindi I</th>
<th>Hindi II</th>
<th>Hindi I+II</th>
<th>Bengali I</th>
<th>Bengali II</th>
<th>Bengali I+II</th>
<th>History</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gayen et al.</td>
<td>.65</td>
<td>.7</td>
<td>.7</td>
<td>.6</td>
<td>.85</td>
<td>.85</td>
<td>.80</td>
<td>.68</td>
<td>.84</td>
<td>.74</td>
<td>.55</td>
<td>.70</td>
<td></td>
</tr>
<tr>
<td>Harper et al.</td>
<td>.80-90</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
</tr>
<tr>
<td>Misra</td>
<td>.50</td>
<td>.76</td>
<td>.63</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.50</td>
</tr>
<tr>
<td>Misra (E.H.)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Reader Reliability

| Taylor et al.       | .75   | (R.A.I) |         |             |                 |          |        |         |           |          |           |           |         |
| Harper              |       |         |         |             |                 |          |        |         |           |          |           |           |         |
| Harper et al.       | .96   | (X)    |         |             |                 |          |        |         |           |          |           |           | .77     |

## Total Reliability

| Gayen et al.        | .50   | .61    | .63     | .44         | .73             | .33      | .51    | .15     | .42       |          |           |           |         |
| Harper              |       | (X)    | (X)     | (X)         | (X)             | (X)      | (X)    | (X)     | (X)       |          |           |           |         |
| S.S.C. Board        | .68-79| .69    | (X)     |             | (X)             |          |        |         |           |          |           |           | .52-53  |
| Misra               | .64   | (P.U.) |         |             |                 |          |        |         |           |          |           |           |         |

Note: "(X)" indicates that the reliability figure is not available.
Contd .... Table showing reliability figures obtained in Indian studies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gayen et al.</td>
<td></td>
<td>.80</td>
<td>.80</td>
<td>.83</td>
<td>.75</td>
<td>.75</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Harper et al.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misra</td>
<td>.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misra</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reader Reliability

Taylor et al.

Harper

Harper et al. .50

(X)

Total Reliability

Gayen et al. .55

(X)

S.S.O. Board,
Maharashtra .40-54

(X)

.49-51 .46-52 .54-62 .57-60 .68-69 .61-70 .71-78 .33-53

.76-80

Harper et al. .50

(X)

Misra
Empirical studies done by the S.S.C. Board, Maharastra (1960), Gayen et al. (1961), and theoretical analysis by Vernon (1940) and Harper (1965) indicate that the total of marks of all the papers of an examination is more reliable than the marks on a single paper.

Regarding the minimum reliability coefficient expected from a good Test. Guilford (1956, p. 146) observes:

Following the leadership of T. L. Kelley, there has been a general tradition that, to be sufficiently reliable for discriminating between individuals, a test should have a reliability coefficient of at least .94. Some have been more liberal in this regard, allowing a minimum of .90, while others have been more demanding, with a requirement of a minimum of .95. While these qualifications mentioned regarding reliability and validity need to be made, the fact remains that in practice we expect reliability coefficients to be in the upper brackets of \( r \) values, usually .80 to .98.

It has already been explained that for practical decision one should consider the total reliability of essay tests. In view of the remark made by Guilford, it is evident that none of the total reliability figures reported in Indian studies reach even the minimum limit of reliability usually acceptable in practice.

IV. Validity

As defined earlier, "a test is valid if it measures well what it is supposed to measure." Statistically, validity is the correlation of a test with a criterion. Studies on the validity of marks are generally confined to the predictive validity of marks for achievement in some other examination.
Tsylor (1962 b) correlated the marks of 120 college students, in Chemistry, Physics, Chemistry Practical and Physics Practical. The coefficient of correlation ($r$) and the standard error of the coefficients ($s$) were found as reported below:

- **Physics Theory vs. Physics Practical**: $r = 0.14$, $s = 0.09$
- **Chemistry Practical vs. Physics Practical**: $r = 0.14$, $s = 0.10$
- **Physics Theory vs. Chemistry Theory**: $r = 0.30$, $s = 0.09$

The last correlation was affected by the presence of two very poor students. If they were excluded, the value of $r$ falls to 0.20, which is barely significant. He concluded that if marks are valid measures of ability, there is no apparent connection between practical and theoretical abilities in Physics, none between the two practical abilities, and none between the two theoretical abilities. As such an assumption is unacceptable, we must conclude that marks are not valid measures at all.

Patel (1962) in an analysis of B.Ed. examination marks found the correlation between the practical annual examination and the year's work in practice teaching to be 0.62. The correlation between total marks in Part I and Part II examination was found to be 0.63. Assuming that the same ability underlies the various papers of the examination, his study suggests moderate validity of examination marks.

Kamat and Deshmukh (1961) inter-correlated the aggregate marks of a group of students obtained in different examinations. They found the following correlations:
These correlations seem to suggest a moderate validity of examination marks. Miara (1964) in a similar study found the following r's.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I.A.</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I.Sc.</td>
<td>.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I.Sc. Ag.</td>
<td>.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduation</td>
<td>.83</td>
<td>.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.A.</td>
<td>.69</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B.Sc.</td>
<td>.85</td>
<td></td>
<td></td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>B.Sc. Ag.</td>
<td>.74</td>
<td></td>
<td></td>
<td></td>
<td>.76</td>
</tr>
</tbody>
</table>

He concluded that the results of high school and intermediate examinations have a high predictive value for achievement in university studies.
Taylor, Tluanga and Misra (1966 c) correlated the marks of 347 students in Matriculation and P.U. examinations. They estimated the correlations between the two examinations making allowance for the students who, because of failing Matriculation, did not appear for the P.U. In other words, their value of $r$ gives an estimate of correlation between the two examinations if all the students who appeared for the Matriculation examinations were allowed for the P.U. examination. The value of $r$ was found to be .72. They also estimated, on the basis of the obtained correlation coefficient, that out of 15,470 students who had failed in the Matriculation examination, 1,688 would have passed if given a chance to appear for the P.U. examination.

Lele et al. (1963) found the $r$ between S.S.C. (English) and P.Sc. (English) marks in M.S. University, Baroda to be .77.

Harper (1963 b) explored the problem of whether the essay or objective tests are better predictors of performance on an essay-type examination. He had an objective examination of knowledge of English poetry as the final "Test Examination" in several Intermediate Colleges. The results were correlated with the U.P. Secondary Boards' examination in the same paper. The validity coefficient ranged from -.25 to .82. (The first coefficient was not significantly different from zero.) In other words, the same objective examination had different validities for predicting the same external essay-examination score in different institutions. The difference between the validity coefficients may have been due to inherent differences between the institutions, or they may have resulted from the extreme unreliability of the essay-type external examination.
Harper (1963 b) in another study administered two objective tests of the knowledge of English usage to 72 Intermediate students. The criteria were two terminal examinations of the essay-type. The correlation between the two essay tests each of 2:30 hour duration was .74. The correlation between the two objective tests each of 45-minute duration was .75. Corrected for the length of the tests, reliability of the objective tests was .91, which is significantly higher than the reliability of the terminal examinations of the essay-type. The validity coefficient of the objective tests, corrected for attenuation, for predicting performance on English essay examination was above +.90. Harper concluded that "the objective test is a better measure per unit time in predicting actual essay writing marks, than an essay examination of the same length." Gyan et al. (1961, 1962 and 1966) have also reported validity of some papers of class X examination of the Secondary Board of Education, West Bengal. To the present research, their validity figures seemed to fit the conditions of reliability and have, therefore, been discussed under the head "Reliability".

The above studies explored validities of marks for predicting achievement in other examinations. However, a more important function of examination marks is to predict performance in actual life. Misra (1964 b) explored this aspect of examinations. He investigated the validity of Teacher Training Courses for predicting performance in teaching on a sample size of 118 male teachers of Boys Intermediate Colleges of Allahabad City. His study suggested that the training of teachers had not improved their teaching efficiency significantly.

It may be observed that an examination is supposed to measure the development of various abilities and skills which education aims to bring
about in students. How far the present examination system measures the development of such abilities is an area almost completely un-explored in this country.

V. Discussion

Before concluding this part of the review it seems desirable to make a note on the practical utility of "validity" and "reliability".

Validity indicates the degree to which a test is capable of achieving certain objectives. There are three types of validity generally recognised, i.e. content, criterion-related and construct. As already stated earlier Indian studies are mostly limited to predictive validity.

Validity is the most important quality of a test. However, validation of a test requires a generally accepted criterion which seldom exists. So in practice a test is considered to be valid, if the questions are relevant to the curriculum and the test is reliable.

As already defined, reliability refers to the consistency of measurement. The consistency of measurement may be considered from two aspects: relative consistency and absolute consistency. (For details see Thorndike, 1949) C

Relative consistency: The coefficient of reliability, which is the correlation between two sets of scores, provides an index of relative consistency. This, being a correlational concept, suffers from the inherent weaknesses of the correlation coefficient. The correlation coefficient
assumes that both the measures are normally distributed, an assumption which is often not satisfied in the case of many educational tests. The farther the deviation is from normality, the more distorted is the value of $r$. Besides, the coefficient of reliability is influenced by various factors other than the intrinsic agreement between the two measures, e.g., the length of the test, consistency of the test items, the method by which the reliability was determined, testing conditions, and the variability in the ability level of the group on which the test was administered. Thus one has always to be on the guard before drawing any inference from the coefficient of reliability.

Absolute consistency: The standard error of measurement gives an index of absolute consistency. It is relatively independent of the group to which the test was given. (See also Harper, 1959 and Misra 1968 b.) Garret (1965, p. 351) holds that the standard error of measurement is "a better way of expressing the reliability of a test than is the reliability coefficient, as it takes into account the variability within the group as well as the self correlation of the test".

However, both the reliability coefficient and the standard error of measurement have limited value for unscaled marks. For unscaled marks, it is quite possible for the reliability coefficient to approach 1.00 and the standard error of measurement to approach 0, and yet on one test all the students may pass and on the other every one may fail. (See also Buros, 1962)

Though nothing can be suggested as a rule of thumb to assess the efficiency of an essay test, as a convenient guide it may be suggested that when two essay tests of the same length are given to the same group the efficiency of the test should be determined on the basis of reliability coefficients.
However, when the tests are of different lengths or are given to different
groups, it may be better to consider the standard errors of measurement.
In each case, the comparison is strictly accurate only when both sets of
marks have the same mean and standard deviation.

3. Suggestions for Improving Efficiency of Examinations

Various ways for improving examinations have been suggested by experts
and studies on Indian examinations.

I. Improvement in Curriculum

Many experts (University Education Commission, 1950; Secondary Educa-
tion Commission, 1952; University Grants Commission, 1961; Gayen et al. 1961;
Examination Committee, 1962; Hill 1964 a, 1964 b; Standards Committee, 1965;
Hill, 1965, 1966 b, 1966 c, University Education Commission, 1966; Hill,
hold that objectives of education and purposes of examinations are not well
co-ordinated in India. Thus, to bring about effective improvement in exami-
nations, curriculum must be improved. Some (Secondary Education Commission,
1952; Gayen et al. 1961; Secondary Education Committee, Assam, 1965;
Hill, 1967 b, 1967 c 1967 f, 1967 g) hold that Secondary Education curricula
should be so improved as to cater to the needs of various types of students:
(a) those who would enter life after passing examination, (b) those who
would enter technical education, and (c) those who would enter higher
education. Hill (1966 c) holds that it appears that the fact that English
is a second language for a great majority of students is over-looked in
framing objectives for the teaching of English. Hill (1966 b, 1966 c,
1967 g, 1967 f and 1967 h) recommends that examinations should measure ability to reason, think, interpret and apply knowledge to new situations rather than the ability to reproduce memorised facts. This means that the curriculum should be clear on what abilities are to be developed by teaching a particular course content.

II. Improvement in Questions

(a) Specificity in questions

Cieslak et al. (1959) recommended that the essay question must carefully define the problem so that students need not guess what the examiner wants; the basis of grading should be clear to examiners and the students.

Ray (1970) recommends that questions should be improved so that they measure the educational objectives set by the curriculum. To be specific, a question should have no room for different interpretations by paper setter, examinees, and examiners as to what should constitute the answer to the question. Specificity in questions has been explored by some studies in India.

Gayen et al. (1961), Lale et al. (1962 a, 1962 b), Hill (1967 f), and Misra (1968 b) found that some of the questions asked in our examinations are vague. Such questions encourage guessing on the part of the students, and lower the reliability of the test. Hill (1967 f) states that in the traditional question papers there is usually no clue, except the number of marks allotted to the question, to indicate the length and scope of the expected answer. Such questions give no stimulus to precision in learning.
Lele et al. (1962a, 1962b) found that there are questions which are considered to be simple by students and difficult by teachers. They concluded that it is probably this differential difficulty of questions as seen by teachers and students that makes essay examinations less reliable.

Some experts have studied the problem of the language of question papers. Gayen et al. (1961) state that "the present system of teaching a subject in one language, but setting of questions in another creates confusion in the candidates; so the media of instruction, of paper setting, and of examination (answering) must be one and the same." The Secondary Education Committee, Assam (1965) issued a questionnaire inviting teachers' opinion on this point, i.e., whether there was any substance in the complaint that the practice of setting questions in English while the medium of instruction was the Regional language was responsible for the large percentage of failure in the examinations. Some of the Headmasters replied in the affirmative. Others did not agree with this view. One Headmaster answered, "A Matriculate or a High School candidate is expected to have at least this much knowledge of English as to understand the meaning of a question set in English Language." It may be observed that however important the ability to comprehend the English language may be, this ability should not be allowed to confuse the evaluation of other abilities. There is absolutely no justification to fail a student in Geography, because he is weak in English.

(b) Increase in the number of questions

The problem of having a limited number of questions in the question paper has been studied by some experts (Harper, 1960, 1962, 1963b, 1963d; Hill, 1965, 1967f, 1967g, 1967h; Misra 1968b). They hold that because
of limited number of questions in the traditional-type question paper, a high chance factor enters in the evaluation of achievement. They recommend the inclusion of more compulsory questions in the question paper, to force a wider coverage of course content. Harper (1960), Hill (1967 f, 1967 g, 1967 h), Misra (1968 b), and Trivedi (1970) recommend the use of short-answer-type questions in the question paper, as they feel that there is no evidence to show that the essential characteristics of essay-type tests cannot be measured by questions requiring answers in one or two sentences, or a page at most.

(c) Reduction in the number of optional questions

One problem which has often engaged the attention of teachers of India and abroad is that whether optional questions should be allowed in examinations or not.

Ebel (1965) recommends that giving choice among questions should be avoided unless special circumstances make such optionals necessary. Other foreign experts (Stalnaker, 1951; Remmers and Gage, 1955; Noll, 1957; Odell, 1958; Wood, 1960; Thorndike and Hagen, 1961; Lindavelli, 1961) hold that if different students answer different questions the basis of comparability of marks is weakened and this would reduce reliability of marks.

There are certain studies in India which have explored the problem of optional questions. Tayler (1962 d and 1963 b) holds that optional questions should not generally be allowed in question papers. Sharma (1963) suggests the provision of choice within a question in such a way that the alternatives do not provide a different type of activity for the students. D'angel (1964)
advocates abolition of optional questions on the following grounds: That the candidate who is expected to answer questions is also required to select them in a hurry. That the process of selection results in loss of time. That under the strain of examination a student may not make wise selection and thus lower down his marks. That the theory of optional questions indicates failure of educationists to agree on what is indispensible to students. That optional questions encourage the practice of not covering the whole course. That optional questions do not allow to have a common standard for comparing students' achievements. That it sometimes results in students misreading instructions and answering more questions or parts thereof than is necessary.

Hill (1963, 1964 a, 1964 b, 1965, 1967 f, 1967 g) holds that if different students answer different questions, no valid comparison of their achievements can be made.

Some Indian studies have analysed the problem of optional questions statistically, on hypothetical cases. Harper (1962) assumed that there are nine questions in a question paper, of which only five are to be attempted. He further assumed that 1/5 of the questions are easy for the student, 3/5 are of average difficulty for him, and 1/5 are difficult for him. He found that there is better than one chance in two that an examination will be "too easy" for a given student. The chances that a student will have to answer two or more of his "most difficult" questions are only three in one thousand. The probability that an examination will be exactly representative of a student's knowledge is only seven out of one thousand. He concluded that the "chance" factor in traditional examinations is far greater than in objective examinations; and, unlike in objective examinations, no statistical correction for guessing in essay-type examinations is possible.
Harper's (1962) paper being highly statistical, Taylor (1962 c) presented Harper's thesis in a simpler way. He assumed, as Harper had done, that a question paper contains nine questions out of which any five are required to be answered. He further assumed that an examination is assumed to test a student's knowledge over a wide field and each item of a test is equivalent and measures a single aspect of knowledge; that a student knows only $\frac{1}{3}$ of the field very well, $\frac{1}{3}$ of the field sketchily, and $\frac{1}{3}$ is not known to him at all; that the maximum mark for each question is 3, a good answer on the average gets 2, and that a poor answer gets 1. He found that in such a case only 21% of the students would get their "true" mark. The performance of the students would generally be very seriously over-estimated. His study suggests increase in the number of questions and reduction in the number of options allowed.

Harper (1963 b) demonstrated with a simple table that, because of provision of choice, a high chance factor enters in the evaluation of achievement in the essay-type tests, and students of the same ability may get different marks depending upon the pattern of choice allowed in the question paper.

The studies done by Harper (1962 and 1963 b) and Taylor (1962 d) based their findings on hypothetical cases. The effect of provision of choice in essay examinations has been studied empirically by Gayen et al. (1961, 1962 and 1966). Gayen et al. (1961) found that an item of 2 optimum marks was balanced against another of 9 optimum marks. One item worth 5 optimum marks was balanced against another worth 13 optimum marks. They define optimum mark of an item "to be directly proportional to its difficulty Value, the
constant of proportionality being the sum total of marks allotted to all the thirty-one question-items of the paper divided by the sum total of the estimated difficulty values of all the items." They further classified choices on the basis of difficulty values of question items. For choice I, the total optimum mark was 89, for choice II the total optimum mark was 110, and for choice III the total optimum mark was 82. They concluded that, simply because of chance choices, a candidate not necessarily of higher calibre gets an advantage over, or fares equally as well as, a candidate of superior merit.

An empirical study by Lele et al. (1962b) supports the findings of Gayen et al. (1961). They found that students avoid difficult questions, and performance on difficult questions is lower than on simple questions. This study suggests that when choice is allowed a student's mark depends more upon his selection of questions than upon his knowledge of subject matter.

Roy (1969) found that some questions have very high frequency of repetition. Due to the provision of choice students generally attempt such "common" questions.

Hill (1967f) reports that an analysis of a Board's question papers showed that because of provision of choice in the Mathematics paper, a student who knew enough about four of the eighteen topics in the syllabus could score the maximum marks. In the General Science paper they could safely ignore Physics altogether. Kaul Adalati (1970) recommends abolition of over-all options.
(d) Use of objective-type questions

Many Indian experts (University Education Commission, 1950; Secondary Education Commission, 1952; Cieslak et al., 1959; Harper, 1960; University Grants Commission, 1961; S.S.C. Board, Maharashtra, 1961; Gayen et al. 1961, 1962; Examination Committee, 1962; Harper, 1962; Taylor, 1962 a; Harper, 1963 a, 1963 b, 1963 d; S.S.C. Board, Maharashtra, 1963 b; Mukherjee, 1964; Hill, 1964 a, 1964 b, 1965; Standards Committee, 1965; Education Commission, 1966; Hill, 1966 b, 1966 c, 1967 a, 1967 b, 1967 f, 1967 g, 1967 h; Harper, 1967 a; Gayen, 1970; Mahajani, 1970; Sreenivasan, 1970) have recommended the introduction of objective-type tests in our examinations. There are several brochures published from the National Council of Educational Research and Training to show how to write good objective-type questions. Despite these recommendations, so far there has been little progress towards the introduction of objective-type tests in our examinations. At the University level they have not been given even a fair trial in India.

III. Improvement in the Accuracy of Marking

Taylor (1963 a), George (1964) and Harper et al. (1967) found that examiners have a tendency to boost up border line cases to the next higher division, which Taylor (1963 a) calls as "J-effect". They hold that the J-effect introduces further errors in marking, and, therefore, suggest the need of impressing upon the examiners not to boost up borderline marks.

Taylor, Tluanga and Misra explored the factors that lower marking reliability of examiners. Taylor and Tluanga (1965) found that there is a positive persistence effect, that is, the following mark is almost always
attracted towards the preceding marks in marking. The occurrence of negative persistence, that is, the following mark being repelled by the preceding marks, is not disproved; but if it occurs it must be very rare. Taylor, Tluanga and Misra (1966 a) found that almost half of examiners are subject to large variations of standards in the course of their marking. The study showed the existence of a diurnal fluctuation in the standards of marking. Taylor, Tluanga and Misra (1966 b), found that examiners' markings are much more accurate early in the process after some twenty-five scripts have been marked. It becomes most inaccurate about two-thirds of the way through. Strict examiners were found to be more accurate in marking. Tendency to mark systematically low or higher was found to be a constant characteristic of an examiner. The tendency to mark on a compressed or wider scale was found to be another constant characteristic. To take care of the above factors in markings, Misra (1968 b) recommended the provision of model answers to examiners, elaborate instructions for marking, and use of question-wise marking rather than script-wise marking. Harper (1963 c) showed statistically that question-wise marking will reduce errors in marks. Gayen et al. (1961) recommended the provision of model answers and elaborate instructions for marking, and giving more time to examiners or fewer numbers of scripts to them. The S.S.C. Board, Maharashtra (1963 b) recommended training of examiners, and increase in time allowed for marking and in the number of scripts re-valued by the Head Examiners.

In certain universities two examiners are appointed to mark the same scripts independently. When the marks given by two examiners differ widely a third examiner is appointed, and the average of the two marks which are closest are given to the student. Taylor and Tluange (1964), in a statistical
study, found that the average of the two marks which are closest may be more erroneous than the average of the two original marks that differed widely. They hold that the improvement in the accuracy of marking when there are three examiners instead of two is only nominal. They recommend that if a script is referred to a third examiner, the average of the three marks should be given to the student. An empirical study done by Taylor, Thulanga and Misra (1966 b) supports the above finding.


IV. Scaling of Marks.

Some Indian studies have explored variability in marking in different papers of an examination, and suggested methods of scaling to make achievements in different papers comparable. Mahalanobis and Chakravarty (1934) found considerable difference in the distribution of marks in various papers of the School Leaving Certificate examination of 1919 of the U.P. Board. They recommended the use of scaling marks of different papers to the same mean and dispersion by the standard score method. Mossain (1940) also emphasised the need of scaling, suggesting the use of a ranking procedure. Bose and Choudhury (1955) recommended the use of percentile scaling. Gayen et al. (1961) found that even the paper-setter does not know the difficulty levels of questions. Their study suggests the need of item scaling before the marks on different items are added. Gayen et al. (1961) recommend the use of equivalent scores before combining marks of different papers. They hold that if scaling cannot be done for all the students this should be done at least for those whose names are to be included in the merit list. Taylor
(1962 e) recommends the use of scaling to equate the marks awarded by examiners differing in standards of marking. Taylor (1963 a) found the use of median scaling easily applicable to large scale examinations. Taylor and Tluanga (1963 b) prepared a scaling Table for scaling raw marks. Harper (1963) recommended the use of non-linear scaling by fixing points at various pass divisions and fail levels. Tluanga et al. (1964) found that the method of scaling as proposed by Taylor (1963 a) legitimately increases pass percentage in examination because it reduces failures due to measurement error. Harper (1963 b) showed that in the U.P. Intermediate examination for a particular year, that in 64 marks of Biology are equivalent to 82 marks in Mathematics, whereas 28 marks of Biology are equivalent to 17 marks in Mathematics. He recommended the use of scaling marks to the same mean and standard deviation before marks of the different papers are compared or added. Hill (1965) also recommended the use of scaling before marks of different papers are compared. In an empirical study, Taylor, Tluanga and Misra (1966 b) found that Taylor's (1963 a) method of scaling reduces error of marking. It may be observed that despite so many recommendations put forth by various experts on examination for scaling marks, no Public examination body in India except the Gauhati University employs any kind of scaling in its examinations.

V. Grace Marks

In almost every Indian examination some grace marks are given to borderline cases so as to enable them to pass the examination. The award of grace marks is often arbitrary. On statistical analysis Taylor and Tluanga (1963 a) found that such a procedure is unscientific. They recommended the use of "passing probability" instead of awarding grace marks. Those who are above a
specific passing probability should be allowed to pass. Assuming standard error of marking to be 5% and 7%, they prepared charts giving calculated values of passing probability for a given raw score. Bora (1963) observed that the use of the passing probability charts should only be made after the standard error of measurement in each paper has been determined by further research.

VI. Change in the Basis of Awarding "Passes"

Gayen et al. (1961) hold that as the difficulty level of a question cannot be determined before students' performance on the question is known, the maximum marks allotted to each question should be given only as tentative in the question paper. The final fixation of maximum marks should be determined after the tabulation of students' performance, in a meeting of paper-setters and examiners. If this procedure is adopted, the chance factor in achievement due to selection of items will be considerably reduced. They also recommend that the minimum marks for getting a division should also be determined each year on the basis of students' actual performance that year.

Hill (1964 b and 1967 c) recommended that for the students who would enter life after passing the Secondary School examination, pass should be the rule and failure an exception. He wanted that we should re-work our standard of pass, and a pass in the Secondary examination should not mean a ticket for admission to university.

VII. Abolition of Supplementary Examinations

Taylor (1962 f), in a statistical study, found that the supplementary examinations enable a large number of undeserving students to pass, whereas the
number of deserving students passing through such examination is only nominal. He recommended the abolition of supplementary examinations.

VIII. Internal Assessment

To improve the reliability of traditional type tests some (Secondary Education Commission, 1952; Members of Education, 1956; Gayen et al. 1961; Flecher, 1961; Vice-Chancellors' Conference, 1962; Corey, 1964; Secondary Education Committee, Assam, 1965; Mahajani, 1970; Mukerji, 1970; Singh, 1970; Roy, 1970; Sarwate, 1970; Chaturvedi and Chaturvedi, 1970; Chatterjee, 1970; Sivayya, 1970; Chavan, 1970; Gayen, 1970) suggest a system of giving some weight to the internal assessment of the students. There are several studies which have tried to explore the efficiency of internal assessment. Pandharipande (1962) found that internal assessment is quite dependable. Jain (1962), Raina (1963 and 1967) and Mallick (1968) found that there is a tendency in the internal assessment to give liberal marks. They also found that in the internal assessment low marks are seldom given. This makes the standard deviation of the internal assessment smaller than that of the external assessment. While Raina (1967) recommended that no weight should be given to internal assessment, Mallick (1968) recommended the use of some statistical methods before internal and external assessment marks are combined.

Some studies (S.S.C. Board, Maharastra, 1960; Gayen et al. 1962; Lele et al., 1962; and Patel, 1962) found that internal and external assessment vary significantly in a majority of the papers. These studies suggest the need for scaling internal and external assessment marks to the same mean and standard deviation before the two are combined.
Misra (1969) found that internal assessment is more consistent from year to year than the external assessment, in awarding marks to students.

Some studies have tried to analyse the effect of combining the internal assessment with the external assessment. The findings of these studies are inconclusive and sometimes controversial. Taylor (1962 a) recommended that the two assessments should not be combined before verifying their accuracies. Without this, it is probable that the total of the two assessments may be more erroneous than a single assessment. He recommended that only the assessment which is more accurate should be considered for giving credit in university examinations. Lade et al. (1962 a) found that combination of internal assessment marks with the external assessment marks does not significantly improve the prediction of future performance of students. Gayen et al. (1962) found that the total of test examination and external examination marks is more reliable than the external examination marks taken separately.

Hill (1965, 1966 a and 1967 h) recommends making internal assessment systematic, and suggests that a student's performance in internal assessment should also be reported along with his achievement in external examination.

IX. Semester System


X. Viva-Voce

4. Conclusions Suggested by Indian Studies

As is evident from the foregoing discussion, various problems of examinations have been covered in Indian studies. However, there is a pressing need for further research. To quote Dave (1966, p. 36):

Although several major dimensions of the field of educational measurement and testing have been tackled ... these dimensions have not yet been exhaustively studied. Much more research work is necessary in practically every aspect of the field in order to create sufficient knowledge, develop new techniques and produce effective material which may ultimately result in the qualitative improvement of educational testing in India.

None the less, the following broad conclusions may be drawn from Indian studies:

1. There is high wastage and stagnation in Indian education.

2. The standards, in terms of passes, of supposedly equivalent examinations differ considerably from one examining body to another in India.

3. The standard of marking varies for different papers of an examination, for different examiners of the same paper, and for the same examiner on different occasions.

4. None of the reliability figures reported in Indian studies reach even the minimum acceptable for grading individuals. Thus, examinations are not fulfilling their avowed purpose to "assess merit".

5. The validity of essay-type tests is very low for predicting job performance, and moderate for predicting achievement in other examinations.
6. There is no extensive study comparing better methods of testing with the traditional essay-type tests. What little research has been done in this line, indicates the superiority of the objective tests over traditional essay-type tests as regards both reliability and validity.

7. No conclusive evidence is accumulated in favour of or against giving credit to internal assessment of the students in the final examinations.

REFORM CONSCIOUSNESS IN INDIA

The studies done in India and elsewhere have had their impact on the thinking of Indian educationists. Several Government of India reports and resolutions have voiced the need for examination reform. Among the early reports, the most conspicuous is the University Education Commission report of 1902. However, during the British regime in India, Indian education was receiving inspiration mainly from the British system of education, where essay-type tests were still dominant. Therefore, nothing concrete was done towards the reform of traditional tests.

The University Education Commission (1950, p. 327) strongly criticised traditional essay-type tests. They observed:

"For nearly half a century, examinations, as they have been functioning, have been recognised as one of the worst features of Indian education in India ... We only note that while the magnitude of the problem has been growing at an alarming rate nothing constructive in the way of reform has happened".
They further observed (p. 328):

"We are convinced that if we are to suggest one single reform in University education it should be that of the examinations.... The essay type of examination ... is ... invalid ... inadequate ... subjective ... and therefore not reliable .... We suggest the introduction of such valid, reliable, adequate and objective examinations in the universities of India at the earliest possible time. Without this there is a danger that Indian higher education will fall into chaos".

The Secondary Education Commission, (1952) recommended the use of a thorough cumulative record on the basis of improved techniques of evaluation. They observed (p. 158): "In order to maintain cumulative records properly, the teacher will have to use a number of tests of different kinds - intelligence tests, aptitude tests and others".

The University Grants Commission's report on Evaluation in Higher Education (1961, p. 21) observed: "Scholastic aptitude tests have been developed to a high degree in many other countries. Such tests could form a valuable supplement to present achievement examinations."

Examination Committee of the University Grants Commission (1962, p. 34) observed: "there are other aims the attainment of which are better evaluated by multiple choice questions, open book examinations, short answer examinations, viva-voce tests, etc."

The Standards Committee of the University Grants Commission (1965, p. 79) observed: "The use of objective tests will not only make internal assessment more reliable, but will also ensure speed, economy and efficiency."
The Education Commission (1966, p. 291) recommended the introduction of objectivity in our examinations. They observed: "Another important point of emphasis would be the reorientation of university teachers to adopt new and improved techniques of evaluation."

It is apparent from these recommendations that there is a great awareness in India of the need for examination reform. Only the desire for change has got to be felt thoroughly enough at the lower levels so that the change becomes possible.

**AGENCIES ACTIVE IN EXAMINATION REFORM**

As a result of the foregoing, many agencies and educational bodies in India have become active in the field of examination reform. Among the most important of these are the Department of Curriculum and Evaluation, N.C.E.R.T.; National Institute of Education (NIE); Evaluation Units of Secondary Boards; Gujrat Research Society; Bureau of Educational Research, Ewing Christian College, Allahabad; Bureau of Psychology, U.P.,; the Kerala, Baroda, Mysore, Gauhati Universities, the Indian Statistical Institute (ISI), Calcutta; and The Regional Colleges of Education. Of agencies working on public examinations, "the Gauhati University among the universities, and the Central Examination Unit are in the vanguard". (Dave, 1966, p. 10)

Besides, several important studies were made under the grant-in-aid schemes of the National Council of Educational Research and Training and the University Grant's Commission. Among the studies done under the grants-in-aid schemes those done by Gayen of Kharagpur, Harper of Allahabad, and Pillai of Kerala are outstanding.
A brief report on the progress of examination reform in India is contained in Hill (1967d).

**PUBLICATIONS PERTINENT TO EXAMINATION REFORM**

Publications in the field of educational research have appeared in great number in recent years, especially from the work of the above-mentioned groups. A few very important ones dealing specifically with problems of examination reform are the following:

Publications of NCERT

*The Concept of Evaluation in Education*

*Improving Examinations*

*Optional Questions in Tests and Examinations*

*Dimensions for Setting Questions in Board's Examination*

*Evaluation in General Science*

*Evaluation in Physics*

*Evaluation in Geography*

*Reports on All India Workshops on the Development of Test Material*

*Booklets on specimen test items in various school subjects*

Besides, *Examination Abstract* Nos. 1 to 6 published by the Department of Evaluation and Curriculum give valuable information about the research studies abroad and in India.
Other publications:

Specimen test items prepared by the Mysore, Maharashtra and other Secondary Boards of Education, and the Punjab University.

Reports on evaluation workshops published by the Mysore, Maharashtra, Madras, Bihar, and Assam Secondary Boards of Education, and the Meerut University.

A Handbook of Selection, Use and Interpretation of Psychological Tests by A. Edwin Harper, Jr., published by U. S. Educational Foundation in India, 12 Hailey Road, New Delhi.

Examination Reforms in Maharaja Sayajirao University of Baroda, published by the M. S. University, Baroda.

PLACE OF THE PRESENT STUDY IN ENCOURAGING EXAMINATION REFORM

From the above observations, we see that almost all the Commissions and Committees appointed to study the problems of your educational system have recommended, in one form or another, the use of objective-type tests in our system of examinations. But there has been little advancement, so far, in this direction. The probable reasons for slow progress seem to be implied in the following statement of the Examination Committee (1962, p.1)

'It is generally admitted that the present examination system in Indian Universities needs reform. The system has been under criticism for a long time. While the need for reforming examinations has been recognised, and much discussion has taken place, little has been done so far to bring about improvements .... There is also a genuine fear in some quarters that the application of
some of the proposed 'remedies' may possibly create a worse situation.

The type of fear expressed in the above report is quite understandable, in view of the fact that improved methods have not been tried in India even on an experimental basis. The following remarks of Dave (1966, p. 13) illustrates this point:

Although several studies have indicated quite clearly the defects of the present system of examinations there are hardly any studies which have undertaken to carry out improved methods on an experimental basis. This is another area where a good deal of developmental research is necessary to improve public examination in India.

BASIS FOR THE DESIGN OF THE PRESENT STUDY

In view of the above-noted considerations, a study was designed to carry out on an experimental basis the use of objective-type tests in Indian conditions, and compare their advantages and limitations with those of the traditional essay-type tests. The following points determined the guidelines for the present study:

1. There has been hardly any study in India to find out to what extent the material covered in essay-type tests in our Public examinations may be converted to objective-type tests. This is the first study on an extensive scale in this direction.

2. There has been no large-scale study in India which ideally satisfies the condition of two independent marking of the same set of scripts for the determination of reader reliability of essay-type tests. In the study done
by Mukerjee (1961) only one script was examined by different examiners. Thus the variability in awarding ranks to the students remained unexplored. It is just possible that two markings may disagree as far as scores are concerned, but agree as far as ranks are concerned. In that case a proper method of scaling may take care of the difference in scores. In the experiments done by George (1964) and Harper (1967), the scripts contained ticks and crosses made by the first examiners, which were likely to invite the attention of the second examiners to these points. Thus the second markings in these experiments do not strictly satisfy the conditions of independent marking. In the study done by Taylor, Tluanga and Misra (1966 b) typed copies of scripts were cyclostyled and examined independently by examiners. Thus, the effect of handwriting, if any, in marking scripts was eliminated from their study.

The strength of the present study is that the scripts did not have any examiners' marks inside them. The number of scripts was large, which would enable us to explore not only the disagreement in marks but also in awarding ranks. The examiners had no clue to know about the assessment given to the answers by the other examiner. Every care was taken to ensure that scripts were not mutilated before they were sent to the next examiner.

3. The present study is the first study in India on total reliability of essay-type tests strictly satisfying the conditions prescribed by Gulliksen (1950) and Harper (1966 a), i.e., the administration of two parallel tests to the same group of students and getting the scripts examined by two independent examiners.
4. Prior to this study, nothing tangible has been reported in India about actually trying out new methods on an extensive scale in the existing examination situation. This is the first Indian study comparing objective and essay tests in a real examination situation such as an examination body must face while introducing objective tests in its examinations. Further details on what "real examination situation" implies will be discussed in Chapter X.