MATHEMATICAL METHODS
APPLICABLE TO THE STANDARDISATION
OF EXAMINATIONS

By

Mary I. Hanania

-

Thesis presented to the Faculty
of the School of Arts and Sciences of the
American University of Beirut
in partial fulfilment of the requirements
for the degree of
Master of Arts in Mathematics

September, 1947.

P r e f a c e

It has once been pointed out that there are some research workers who use statistics as a drunken man uses lampposts - for support rather than illumination. One of the aims of this study is to reveal ways in which the research worker in education can adopt statistical methods, without necessarily falling under that category.

A large portion of the paper is more in the nature of a survey than a concentrated study of any one particular problem. Three independent enquiries described in the last chapters serve as hints on some possible applications of statistical tools and methods. The rest of the work is mainly compilation and classification of methods, the testing of some of them, and a brief discussion of their values.

In the testing of the methods, all illustrations - with one or two exceptions - were taken from mathematics examinations held at this University and in the Palestine Matriculation. In this connection, I wish to thank the Dean of the Board of Higher Studies in Jerusalem for kindly allowing the use and publication of their data and for reading and commenting on the draft for chapter V. I am also grateful to Professor Mulholland of this University, who supplied the data from Freshman mathematics examinations, and who supervised my work.

September, 1947. M. Hanania.

Table of Contents

Introduction

Since its main problem is the employment of mathematical methods in improving examinations, this thesis is assuming the validity of the present educational systems that fall under discussion and that require the holding of such examinations. Rather than speculating on purely educational issues that arise out of a study of examinations and their function in any institution, it has for its purpose the exposition of the nature and use of various statistical methods that are applicable to standardising those examinations, against a bare background of educational facts.

The main body of the thesis lies in chapter III, which is a straightforward description of the methods that have been or may be applied to the problem. This description rests largely on the theoretical discussion of chapter II, and draws, in certain cases, on material surveyed in the first chapter. The essential statistical constants and symbols, in uniform use throughout the paper, are introduced in the third chapter and, although there is frequent reference to them in the earlier chapters, their function there is wholly descriptive.

Since some of the terms used in educational statistics are likely to cause misunderstanding in interpreting them, it will be better to define them as they are used here, together with a few statistical terms of frequent occurence in the coming discussion :

A <u>score</u> is the primary numerical evaluation of part or the whole of a piece of work.

A <u>grade</u> is a category of work used in classifying achievement (general, or on a particular examination), the letter grades being those most commonly used.

A <u>mark</u> is an expression of the evaluation of his work given to a student or candidate either in numerical form (usually a percentage) or as a grade. A score, therefore, may be used as a mark or converted to one on the proper scale.

The <u>reliability</u> of a measure is taken as its consistency in measurement, and the <u>validity</u> of an examination as the degree to which it measures what it purports to measure.

<u>Correlation</u> between two sets of marks is a measure of the degree of correspondence between them.

An examination is said to be diagnostic when it differentiates sharply between students of diffferent abilities.

A new-type examination, in contrast with the essay type, consists of a large number of short questions requiring brief answers, and scored objectively by merely counting the number of correct or incorrect answers.

If, in a number of measures that contribute to a total, some are weighted, or given special relative values, their contributions to the total are proportional to their respective weights.

A certain specified group of individuals subjected to statistical study is called a population and any random subgroup of this is a sample of the population. The constants of the distribution in the population are parameters, their estimates from samples being termed statistics.

The normal distribution curve is the bell-shaped curve, symmetrical about the Y axis, whose ordinate at any x is defined as

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

A test of significance distinguishes two classes of results in a statistical study :

(a) one which shows a significant discrepancy from a certain hypothesis,

(b) one which does not show such discrepancy.

The hypothesis tested is called the null hypothesis which is liable only to be disproved - if the result is of type (a). Levels of significance are usually taken at 5% and 1% which imply, respectively, a probability of 5 and 1 in a hundred that the hypothesis holds in the case subjected to study. A discrepancy with a probability of occurence in random of less than 5% is usually taken as significant.

There are so many different aspects of the problem studied in this paper that it was found impossible to draw any conclusion which can adequately sum up the results. Instead, the final chapter has been dedicated to the planning of a proposed scheme in standardising examination marks - utilising some of the methods of earlier chapters; and, although part of the plan has been tested on a small group and found effective, conclusive evidence based on a larger investigation is still lacking.

Chapter I

A SURVEY OF AVAILABLE LITERATURE ON STANDARDISING EXAMINATIONS

Within the limited scope of available books related to the subject, some educationists were found to treat the problem of unstandardised examinations as purely educational both in nature and remedy; there were others who, with the introduction of statistics in psychology, sought work on a scientific basis by the application of elementary statistical techniques and notions in standardisation; and still others preferred to start with the theory behind the subject, using methods of higher mathematics in the analysis. Yet all had one common point of view - that of a dissatisfaction with current examining and marking systems and a recognition of the need for better standardisation, because of the importance of examinations in evaluation and academic achievement and their bearing on guidance for future careers.

The following survey, while introducing separately the books and articles with a brief reference to the contents of each, will serve as a definition of the scope of the subject under study and of what has already been accomplished in it.

In his book, Examinations and their Substitutes in the United States, Kandel[1] enumerates the three central problems of education as:-[2]

1. The setting up of generally accepted standards of achievement;
2. The devising of methods of measuring this achievement, and of holding pupils to performance;
3. The introduction of such flexibility in educational offerings that each individual may receive the education from which he is able to receive the greatest benefit.

His second problem is evidently that of examinations.

He reviews the movement for standardising those "measuring devices", Starch and Elliot's investigations on the reliability of examinations, the establishment of standardised tests by Thorndike and others, and the introduction of the new-type objective examinations. He criticises the latter on the basis that they test only memory of

---

(1) In this chapter, the titles of books reviewed and the authors' names are underlined at their introduction.
(2) see his Preface.

facts and are therefore not applicable to all subjects, and he recommends that the marking of essay examinations be made more reliable by formulating more specific and restricted types of questions.

A study of the "Use of Psychological Tests in England" is given by Burt in the International Institute Examinations Enquiry's Essays on Examinations [1]. He points out as the chief difference between a psychological test and a school examination that, whereas the psychologist has worked beforehand to standardise both the methods and the results, the main work of the ordinary examiner takes place after the examination is over. As will appear later, this is not the case in a good examination, and P.B. Ballard, in an essay [2] on "The Special Place Examination [3] ", criticises the examiners who employ an unreliable mode of examination and apply trustworthy tests to borderline cases only. "The examiners in question", he says, "first measure with a barge pole and then worry over millimetres".

In a note on "The Reliability and Validity of Measurements" [4], Spearman defines the former as "the degree of agreement between any two independent sets of measurements of the same set of things", and the latter as "agreement of measurements with the things measured". It follows that low reliability involves low validity, while the converse is not necessarily true.

Traxler enumerates five "Problems arising out of the attempts to apply improved measurement techniques to education and guidance" [5] . The problems are:-

1. Selection of test items;
2. Comparability among tests;
3. Diagnostic measurements;
4. Relationship of factor scores to future educational and vocational success;
5. Scoring objective tests.

He suggests that the first problem be met by the examiner's stating his objectives before selection of the items, and the second by establishing common national standards, or fixed scales.

---

(1) p. 98 -- .
(2) Essays on Examinations, pp. 116-117.
(3) A competitive examination held by local educational authorities in Britain, on the results of which pupils from elementary schools are either a) awarded scholarships for public schools,
b) admitted to "central schools" up to 15 years of age, or
c) retained until school-leaving age.
(The Marks of Examiners, p. 68)
(4) Essays on Examinations, pp. 108-9.
(5) Journal of Educational Research, September 1943, pp. 14-18.

Monroe also recognises the importance of selecting the right items in an examination. Comparing "Educational Measurement in 1920 and 1945" [1], he finds that the major question in 1945 is the validity of the examination relative to a specified purpose. There is now need, he says, for explicit measurement of all aspects of educative growth.

Gregory gives in The Fundamentals of Educational Measurement a summary of types of examinations, methods and means of examining. He classifies standards of examining into -

a) Standards of quantity,
b) Standards of time, and
c) Standards of quality.

Scoring may be done in two ways:-

a) By accumulation scores, where each question (or part of question) is given a weight and the sum of the scores constitutes the final mark;
b) Scoring by greatest difficulty, where the final mark is determined by the weighted value of the most difficult problem a pupil can solve.[2]

His second method is sometimes applied to mathematics examinations, but it cannot be reliable due to the chance factors involved. It may, however, be used with advantage in selective tests where the paper has a wide and carefully graded range of difficulty.

Vernon holds that the main requisites of good marking are a similar distribution and a similar average level of marks in all subjects. In The Measurement of Abilities he surveys various methods of marking (ranking, grading, numerical evaluation, etc..) and suggests that all marks be reduced to some special scale [3] (for comparison) by the percentile technique [4]. He deals fully with elementary statistical procedures in the measurement of abilities, and dedicates a fair portion of the book to a discussion on mental tests and objective type examinations and their construction. One of his application of statistics to mental tests is the factor analysis of human abilities into a general factor, group factors, and specific factors detectable in every test [5]. Many of his hints and suggestions on marking and on setting examination papers should prove to be a valuable guide to examiners.

---

(1) Journal of Educational Research, January 1945, pp. 38-42.
(2) for difficulty of a problem see below: Ch III C. para 1, Ch. V.
(3) cf. Traxler, p. 4 above.
(4) explained below: Ch. III C. para 4.
(5) cf. Ch. II below, on "Factors".

Ruch sums up, in The Improvement of the Written Examination, what is accepted by all - though not everywhere explicitly stated - to be criteria of a good examination [1]. A good examination, he says, must be:

1. Valid,
2. Reliable,
3. Objective - or free from the personal element of judgment in the scoring of answers,
4. Easy to prepare and score,
5. Conforming to certain specific standards.

After a review of studies on types of examinations he recommends the objective or new-type examination as the best available example of a "good" examination. More recent investigations, however, still accepting his criteria, find the use of standardising methods on all types of examinations a better policy.

One of the wide, and comparatively recent, investigations on reliability is Valentine's The Reliability of Examinations - "an extensive enquiry as to the reliability of examinations as tested by subsequent performance, with special reference to entrance examinations to secondary schools and to the awards of University scholarships". In five out of the eight centres studied there was practically no relation between the order of merit in the entrance examination and the order of merit at the end of the secondary school carrer [2], - the average coefficient of correlation [3] between the two orders being practically 0.

Discussing the causes of this unreliability, Valentine suggests that the element of 'luck' in answering be partially met by avoiding too narrow a range of questions; that varying 'form' of candidates be met by the offer of a possibility of a make-up, at least for borderline cases; and that the examiners' standards be made to compare more favourably with each other [4].

Monroe gives a detailed study of the theory underlying unreliable examination marks. He describes his work on The Constant and Variable Errors of Educational Measurement as an aid to a more intelligent use of educational tests. He analyses any obtained score on a test or examination into three additive components [5] :-

Obtained Score = True Score + Constant Error + Variable Errors,

where a constant error is one which has the same magnitude and sign for all of the scores of a given group,
a variable error one which varies in magnitude and sign for several scores of a given group.

---

(1) Ruch, Ch. II.
(2) Valentine, p. 59.
(3) for computation see below, Ch. III C para 5,
0 correlation indicates absence of correspondence.
(4) Statistical methods for standardising between examiners will be dealt with in Ch. III.
(5) This property of the components is assumed primarily for simplicity of computation, such assumption being permissible (Burt - Marks of Examiners p. 251) in first approximations.

According to the theory of errors, approximately half the variable errors for a given group of scores (if the group is large enough) are positive and half negative, since they are random errors, and if assembled for a frequency distribution the shape would approximate the normal probability curve with the average at zero.

A similar analysis is expounded by Rhodes in The Marks of Examiners, where he considers the "ideal" or standard mark for a unit piece of work, and describes the obtained score in terms of this ideal score. Taking a unit piece, an examiner will assign to it a mark A + d + e, where A is a measure of his general level of marking, d the deviation by which he estimates the candidate to vary from the general level, and e a measure of his relative error. His "ideal" score would therefore correspond to Monroe's "true" score,

A is the true average + constant error of the examiner,
d the true mark - the true average,
e the variable error of the examiner.

By making a number of examiners correct the same scripts, Rhodes devised a method for reducing those constant and variable errors in each candidate's marks - thus approximating them to the "ideal" marks:

1) He found the simple average of each candidate's marks by adding the marks given to him by all the examiners (n in number) and dividing by n).
2) He calculated the differences d_{ik} of each examiner i's marks from the average of candidate k in 1).
3) The constant error CE_i for examiner i is the average $\sum_{k=1}^{m} d_{ik}/m$ of his differences for all the candidates, m in number.
4) He expressed each d as the sum of consistent and variable errors, i.e. d = CE + VE.
5) He calculated the variance[1] of random (or variable) errors for each examiner i, or s^2_{VE}, and considered its reciprocal $1/s^2_{VE}$ as weight w_i.
6) The "ideal" mark then for each candidate was found as $\sum w_i M_i / \sum w_i$, where M_i is the mark given to him by examiner i.

The Marks of Examiners is the result of investigations carried out in England by Hartog and Rhodes after an International Conference on Examinations held in 1931. They obtained written scripts in a number of subjects in differing fields and asked different examiners to mark them, using different combinations of examiners and different examining conditions - with the object of arriving at conclusions regarding discrepancies in examiners' marks. Following is a summary of their results [2] bringing out clearly the outstanding errors in marking.

---

(1) see Ch. III for definition and mode of calculation.
(2) on "The Marks of Examiners", Journal of the Royal Statistical Society, Vol. C. Part I, 1937, pp. 106 - 110.

Examination	a	b	c%	d%	e%	Remarks
School Certificate History	15	15	16	6	5	Scripts of the same original middling mark; no instructions; remarking after one year.
School Certificate Latin	15	15	16	4	2	Scripts of exactly the same original mark; maximum 350.
School Certificate French	50	12	8	16	2.5	Scripts with original marks roughly normally distributed; detailed marking instructions.
School Certificate Chemistry	30	12	17	19	4	Originally normally distributed.
School Certificate English	48	7	15	6	4	
Special Place Essay: Impression	75	10	28	(g)	8.4	
Special Place Essay: Details	75	10	13	(g)	7.9	
Special Place Awards	150	10	-	-	-	
Mathematics, Honours	23	6	3.5	20	4	6 out of 12 questions.
Entrance School Essay	50	5	4	10	7	1 of four topics.

Notes:

a - Number of scripts selected for investigation.

b - Number of examiners (experienced in type of examination selected).

c - Range between average marks awarded to the a scripts by the b examiners.

d - Average standard deviation of marks of any one examiner,
Mean deviation of marks of any one examiner, or
Standard deviation of ideal marks for all candidates.

e - Standard deviation of examiners' random variations

(g) Average of standard deviations of examiners' distributions = 16.0, 16.0.
Standard deviation of ideal marks = 14.4, 13.9.

In a memorandum to this book on "The Analysis of Examination Marks", Burt shows that factor analysis [1] may be used to determine a candidate's hypothetical true mark. Assuming that the marks of any given examiner may be resolved into two hypothetical components -

(1) the true value - a component influencing all examiners in different degrees, and
(2) the residual errors - a component peculiar to each examiner,

one can calculate how closely the marks of that examiner approximate to the hypothetical true marks and how far he is influenced by irrelevant factors of various kinds. "Where the correlations between the several examiners and the true mark are not likely to differ widely, the unweighted average of the marks allotted by all the examiners yields a fair and quick estimate of the ideal or true mark.....Where the correlations between the several examiners and the true marks differ widely, the marking of the best [2] examiner is almost as accurate as the average marking of the whole board, and may even be more accurate." [3]

Another investigation on reliability of written examinations was carried on a smaller scale by Monroe and reported in Educational Tests and Measurements. The procedures adopted were similar to those used by Hartog and Rhodes. Two examinations were given to the same pupils, in most cases the questions being prepared and the papers marked by different examiners, and the correlation between the pairs of sets of scores measured. The results varied widely from a very high reliability coefficient [4] to a negative one.[5]

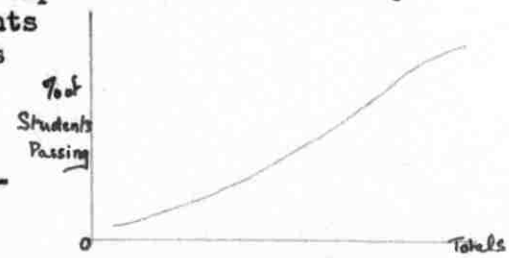
The book gives examples of standardised tests, which are expected to possess a higher degree of reliability than ordinary written examinations. Monroe's standardised reasoning test [6] in arithmetic follows a scheme of marking for solutions correct in principle and for correct answers [7], which is useful in the marking of ordinary mathematics problems.

Paterson's suggestions for increasing reliability of examinations are by the use of objective examinations to replace the common essay type. He gives full directions for constructing them in his Preparation and Use of New Type Examinations. Among his directions is a suggestion for compiling a reservoir of "diagnostic questions" [8] for

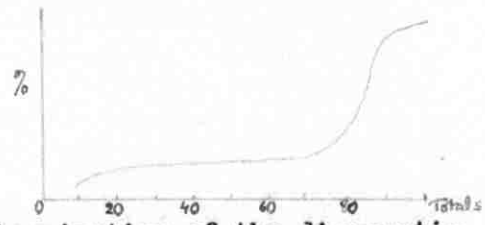
---

(1) see Ch. III C below.
(2) as defined by the highest correlation with the true mark.
(3) The Marks of Examiners, p. 310.
(4) cf. Ch. II below; interpreted like a correlation coefficient.
(5) Monroe, De Voss and Kelley, p. 470.
(6) Ibid. p. 62-.
(7) cf. Ch. II below.
(8) Paterson, p. 60.

ready assembly of highly accurate examinations, by analysing the results of each question after an examination (where the number of students is sufficiently large) and discussing its "diagnostic curve". The curve is drawn by plotting the percentage of students passing the question against the totals obtained by the students on the whole examination. For an ordinary examination question the curve is expected to rise steadily - showing an increase in the number of students passing the question in the better sections of the class.



A question whose curve makes an abrupt rise is a good diagnostic question - because it differentiates rather sharply between students of different categories of achievement - and its degree of difficulty depends on how far along the "totals" scale this change occurs. For instance, a first-class good diagnostic question may be represented by a curve like this, in which it was mostly students with a total greater then 80 who were able to pass the question.



This method is applicable to the determination of the diagnostic significance of any question in an examination, as is illustrated in a later chapter [1].

Another advocate of the new-type objective examination is Wood. In Measurement in Higher Education he gives a full description of Thorndike's Intelligence Examination, pointing out its advantages, and giving suggestions based on it for constructing new-type examinations.

Peters and Martz carried out a "Study of the Validity of Various Types of Examinations" [2]. An examination was given to all classes beyond the second, and was set up in four forms:-

(1) True-False statements.
(2) Four-alternative multiple choice items.
(3) Completion exercises.
(4) Essay discussion type (graded according to uniform rules).

The criterion of validity was the final grade assigned by the teachers (based upon the aggregate of the objective tests and upon the teachers' estimates of pupils' performance). The results of the four types were then compared with this criterion. They concluded that the four types did not vary greatly in validity, and that type (4), when objectively scored by fixed standards, was valid in any grade.

---

(1) Ch. III below, section 1.
(2) School and Society, Vol. 27.

Sims experimented on "Improving the Measuring Qualities of an Essay Examination"[1] and found that it was possible to reduce disagreement by:-

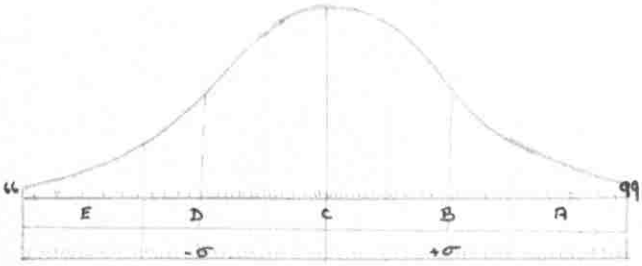
(a) the use of certain scoring rules, and
(b) converting raw scores into grades - thus reducing constant errors of measurement.

This latter device is merely a linear transformation of the scores to a fixed level and range[2].

Stevason adopts a non-linear transformation of scores to school marks, based on the theory of an expected normal distribution of marks in the class. In "Simplifying the School Marking Process"[3] he describes his "Marking Scale" and explains its use. It can be constructed by the use of statistical tables for a normal distribution, and contains four scales (to adapt to any system of marking):-

(1) Percentage marks (66 - 99).
(2) A grades scale (A, B, ... on the 10-20-40-20-10 frequency scheme).
(3) % distribution, reading both sides from the mean.
(4) Standard deviation points - reading both ways from the mean.

The marking of an interval on the percentage marks scale is proportional to the area under the curve above that particular interval.



Since this method causes scores to accumulate about the centre, if the individual scores of separate teats are first referred to the normal curve by Stevason's scale, and then added, the final averages will show an abnormal tendency towards the centre. In this case it may be better to convert the totals by this method and leave the separate items with unaltered scores.

Another method of conversion is used by McCall and Bixler in How to Classify Pupils. Here they change raw scores into "grade scores" which express the achievement of a pupil in terms of the ability or achievement of the average pupil of a given grade throughout the nation.

---

(1) Journal of Educational Research, September, 1933.
(2) described in Ch. III C below, section 4.
(3) Journal of Educational Research, April 1945.

The method, applicable only to standardised tests, lies essentially in giving a very large number of pupils of specified grades a certain standardised test and averaging their scores for each grade. Then each grade is plotted against its average score on an XY plane and a smooth curve run through the plotted points. An intermediate grade score can then be readily located on the graph for any pupil who takes the test and obtains a known score on it.

In The Improvement of College Marking Systems, Spence describes a procedure for combining a student's marks into one figure in order to get the best measure of his total work. The advantage in this is that in combining independent measures each in itself of small reliability into a total measure, the variable errors tend to balance and the result is of greater reliability.

The procedure rests on calculating McCall's T-score[5] for every individual student in the college. McCall's T-scale is again based on the assumption of a normal curve of distribution. Weights are assigned to the various marks on the basis of their standard deviation positions. If d is the standard deviation position of the mark on the scale, $D = 5+d$ its position referred to -5σ as origin, then the T-score is defined as $10 \times D = 10(5+d)$.

Spence's proposed plan for handling grades rests on the use of T-scores. Every student's T-scores (in all subjects) are computed and averaged in the office after the teacher has assigned ranks to them[1]. The students can then be compared on a corrected basis.

Then, by constructing frequency distributions[2] of T-scores in the various classes, the mean, median and standard deviation[4] of T-scores for each class can be computed, and the classes thus compared as to level and variability on a uniform basis.

One of the main disadvantages of such plans as T-scaling is the amount of office work involved after the teachers hand in the marks - while it is important to improve first every individual' examiner's marking system rather than concentrate on office work.

Soderquist devised a "New Method for Weighting Scores in a True-False Test"[5] which calls for the weighting of his responses by the examinee himself according to the degree of assurance with which he makes the responses[6]. He tested the method in 75 items on 62 candidates, and concluded that weighting for assurance in a test had the same effect

---

(1) cf. Ch III, section 4.
(2) Spence, p. 64 --.
(3) & (4) cf. Ch. III below.
(5) Journal of Educational Research, December 1936.
(6) for details of method employed cf. Ch. III C, section 3.

as lengthening such a test [1], and that there was less guessing under that method than usual in true-false tests.

Odell carried out an investigation on The Use of Scales for Rating Pupils Answers to Thought Questions. He built up a scale for each question to serve as a basis for comparison of answers. To build up the scale he collected for each question a number of answers and selected papers which, according to the average judgment of several raters, deserved 0, 1, 2, ... 10 rating. To these he added any criticisms presented. Then each question, with the set of 11 answers and criticisms, formed a "scale".

The investigation, however, led to no positive results, as rating with scales did not prove significantly more reliable than without them.

In an article on "Previous Class Cumulative Index as a Guide to Grading"[2] Leker describes another plan [3] for guiding the examiner in assigning marks - a plan based on the assumption that the general ability of a class, expressed in terms of the median score and quartile deviations [4], is stable (within a narrow range of variability) from year to year. The cumulative index of a class is defined as the average of the cumulatives of each student in the class, expressed in index form.[5]

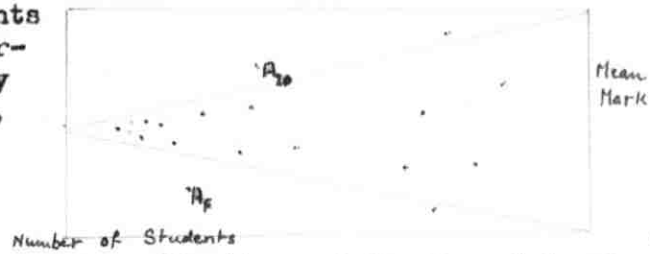
A large scale statistical study for comparing general levels and scatter in the marking of large examinations by a number of examiners was reported by Sandon in an article on "Control Charts in Script Assessment in Large Written Examinations".[6] A secondary school admission examination was given in England to 3000 children, 1500 of whom were boys, in 23 centres. The examination consisted of two arithmetic and two English papers, and each paper was marked by one of three sub-examiners. The results were then analysed by Analysis of Variance[7], and then by the plotting of Control Charts.

A control chart is plotted on a pair of rectangular axes, the vertical representing the mean mark in %, and the horizontal the number of students. Assuming a normal distribution of marks for each paper, two "control lines" are drawn within which most of the averages are expected to fall [8]. Then the average grade for every centre and every examiner is plotted on the chart, and the position of this point with

---

(1) and therefore increasing its reliability (Ch. II below).
(2) Journal of Educational Research, September, 1945.
(3) Ch. III, section 3.
(4) constants of a frequency distribution, described in Ch. III B.
(5) The cumulative of a student for a period of study is the sum of his marks in all subjects for that period; an index is the % ratio of a value obtained to a fixed standard value.
(6) Journal of the Royal Statistical Society, Vol 54, pp. 343-8.
(7) Ch. III C below, section 5.
(8) mathematical basis for construction in Ch. III.

respect to the lines can be interpreted as a deviation. If point A_{20}, for instance, represents the group (centre 20) examined by A, we can conclude that A overmarks students in that centre, whereas he undermarks girls' papers - judging by the position of A_F on the graph.



A detailed analysis of the applications of the Normal Curve to examinations is given by Rugg in Statistical Methods Applied to Education. He emphasizes its use in anticipating distributions of examination results, and describes the method of evaluating the difficulty of examination questions by an application of this curve[1].

In Psychometric Methods Guilford describes the validity of a test by the discrimination value of its items - which he determines by curves similar to Paterson's diagnostic curves[2], where the proportion of success is plotted against scale of ability (as judged by totals on the examination or test, probably). Then the median of the curve represents the level of difficulty of that particular item, while the slope is proportional to its diagnostic value.

Guilford deals, in particular, with the difficulty of test items, and refers especially to psychological tests. Since most of the methods employed in standardising school examinations have been borrowed from the field of psychology, it is not surprising to find in the discussions of so many writers school examinations converted to simple objective tests. Methods for standardising tests (like Guilford's method) may prove to be of special value if adapted to ordinary examinations, but the problem of examinations comprises a far wider field and is affected by other additional factors.

---

(1) Ch. III c , section 1.
(2) cf. p. 10 above.

Chapter II

FACTORS IN EXAMINING AND MARKING

(based on results of previous investigations)

If it was one day decided that examinations be abolished, few students or teachers would not accept the news with great satisfaction. Nevertheless, much to everybody's discomfort, examinations cannot be abolished from our present educational systems - because of their immense guiding value to the educator. For, besides their indirectly serving as incentives to students in schools and colleges, the main function of examinations is the evaluation of the amount of required knowledge a student has absorbed and retained. Moreover, they set educational standards for selective purposes, as in admission to college classes, in placement examinations, the award of scholarships, or in employment.

Accuracy in examining is therefore not only desirable but essential - and the examiner should aim at as much perfection in setting and marking the examination as the most scrupulous student in preparing for it.

An examination may take one of a wide range of forms adaptable to different subjects or different examining conditions and with varying degrees of liability to standardisation. The main types are:-

1. Practical or Performance Examinations:
   These are comparatively infrequent in use and, once the objectives of an examination are stated clearly before the performance, they require little more than the evaluation of a "pass" or "fail".

2. Oral Examinations:
   These have the advantage of testing the whole range of knowledge; but so many psychological and chance factors are involved in the process, that it is very unsafe to judge a candidate by such an examination alone.

3. Written examinations, comprising objective questions, essay type questions, and problems, are the most widely used, and it is on them that practically all work in standardisation has been carried.

While some methods (modified, if necessary) may with advantage be applied to practical and oral examinations, discussion in this paper is confined to the third type and all illustrations drawn from it. Standardised tests, written and oral, are excluded.

Of the five criteria for a "good" examination listed by Ruch, reliability lends itself most to statistical description and measurement; it is therefore a suitable means for objective study, and a numerical measure of the reliability of an examination is a valuable guide in its standardisation. If two examiners mark the same set of scripts, the two sets of marks are expected to correlate highly - provided, of course, that the markings are "reliable". Similarly, a low correlation would indicate low "reliability" of marking. Thus the correlation coefficient r [1] between thhe two mark sheets may be taken as a measure of the reliability of the marking - or, simply, as the "reliability coefficient". In the same way, if duplicate [2] examination papers are given to the same class, the coefficient of correlation between the two sets of scores is the reliability coefficient of the examination. Or, in one single examination with a large number of short questions of the same difficulty, self-correlation (or the correlation between scores on alternative questions) is taken as the reliability measure.

Like the correlation coefficient, the coefficient of reliability ranges in value between -1 and +1, and the significance of each value obtained may be read from tables (Fisher and Yates, Table VI). A value of the coefficient smaller that .9 is, however, an indication of low reliability - since a truly "reliable" examination should correlate perfectly with its duplicate, only a small range of error being allowed for differing examining conditions.

To a reliability coefficient of .9 corresponds a coefficient of alienation [3]

$$k = \sqrt{1-.81}$$
$$= .44 \quad \text{(approximately)},$$

which indicates the degree of lack of correlation.

A reliability coefficient of .7 between two markings of the same paper shows a definite tendency to correlation, but also implies an equal degree of non-correlation.

In the literature of educational research there is abundant evidence on the unreliability of the usual examining and marking systems. Among others, Monroe, Valentine, Hartog and Rhodes report very low reliability coefficients obtained in independent studies. In a series of examinations, for instance, in which two teachers marked the same class every time, the median [4] reliability coefficient obtained by Starch and Elliot was .65 [5] - a very low value.

---

(1)& (3) Ch. III B.
(2) identical in type and, as far as possible, in difficulty, but differing in specific content.
(4) or middle measure. Ch. III B.
(5) Monroe, De Voss & Kelley, p. 470.

On the other hand, it was found that the reliability of an examination could be increased by adopting certain devices in the setting of questions - such as in stating the objectives of each question beforehand, expressing questions in specific or standard form, employing a scheme for marking [1], and increasing the number of questions thus allowing for a wider scope of the field. That a longer examination can be more reliable than a short one (other factors being kept constant) may be demonstrated by reference to the Spearman-Brown Prophecy formula which expresses the correlation coefficient derived from a multiple of a set of scores in terms of the correlation coefficient derived from one set only as - [2]

$$r_N = \frac{Nr}{1+(N-1)r},$$

where r_N is the reliability coefficient of the longer examination,
N is the number of times it is longer,
r is the reliability coefficient of the short examination.

This apparent increase in reliability led to a more extensive use of objective examinations, characterised by such properties as those mentioned above [3]. In fact, wherever applicable, these examinations have proved very useful, though not perfectly reliable, and certain methods in their marking have been devised for better standardisation [4].

In analysing examination marks and the inaccuracy involved in them, the best and simplest method of approach is that taken by Monroe in his discussion or errors in educational measurement [5]. Ideally, we may regard the performance of a student in any particular examination as deserving a certain mark (T) - the "true" or "ideal" mark that should be objectively given to that class of work. In practice, an examiner's evaluation of that particular piece of work may deviate by a considerable amount in either direction, becoming ($T \pm d$), and the problem then is to standardise the examination so as to reduce (d) to a minimum and obtain the nearest approximation to the "ideal".

---

(1) In a Special Place English examination, Hartog & Rhodes (The Marks of Examiners) compared the discrepancies between the marks awarded by ten examiners when the essays were marked on impression only with the discrepancies which occured when they marked in accordance with a detailed marking scheme. Each examiner marked 75 papers on impression and 75 others (from the same homogeneous set) according to the scheme. Then his marks in both cases were added and averaged. The range betweem examiners' averages was 28 in a hundred for marking on impression and 13 for marking by the scheme.
(2) Appendix A, para 1.
(3) Kandel, p. 85; Ruch, p. 121; Vernon, Ch. XII.
(4) Chapter III C, section 3.
(5) cf. Ch. I above, pp. 6-7.

This deviation (d) may be the result of several factors in examining and marking conditions - such as differences in the standards of examiners ~~while~~ and various influences that may affect the examiner while marking any particular paper. Discrepancies in marks due to the latter constitute "variable errors" - (v), so called because they vary between different individuals in the same group and subjected to the same test. If duplicate tests are given to the same class and corrected by the same examiner, variations in the marks of the same individuals are an indication of variable errors of marking, and these errors can therefore be evaluated by calculating the coefficient of reliability of the tests.

The other error in marking (c=d-v) which, together with variable errors, constitutes the total discrepancy in an examiner's marks, is the "constant error" for each examiner - being due to differences in subjective standards of examiners (their degrees of severity) and to the degree of difficulty of the examination paper.

Every examiner's deviation from the true mark constitutes both constant and variable errors which cannot be readily determined, and a ~~reductio~~ reduction of constant error by a linear transformation of the whole set of scores given by any examiner does not suffice to eliminate all discrepancies due to him - for one might easily conceive of an examiner who not only fixes his average at a certain eccentric level but also allows only very few markings to deviate freely from that level; in other words, besides introducing an error in the mean of his marks, his marks have to be corrected for too low a standard deviation [1]. A comparison of the marks given for the same set of scripts by different examiners can give an indication of the differences in standards between them and the amount of variation in the marking of each examiner by himself.

Given an array of the marks of m examiners for n candidates, Rhodes [2] obtained an expression for calculating every examiner's standard deviation from the "ideal" mark:

If X_j is the mark awarded by A to candidate j,
then $X_j = \overline{X} + x_j$, where $\overline{X}$ is the examiner's average mark,
x_j his deviation for j from the average.

Similarly, Y_j, Z_j, etc... are marks awarded by B, C, etc... to the same candidate. Then A's standard deviation is given as -

$$s_\alpha = \sqrt{\alpha/n} \quad [3],$$

where $\alpha = m/(m-2)\sum x_j^2 - 2/m-2 \sum x_j(y_j + x_j + z_j + \ldots) - 1/(m-1)(m-2)\left[\sum x_j^2 + \sum y_j^2 \ldots\right] + 1/(m-1)(m-2)\sum(x_j + y_j + \ldots\ldots)^2$,

the summations being over the n values of j.

---

(1) for a mathematical definition of these terms see Ch. III B below.
(2) The Marks of Examiners, p. 193.
(3) Appendix A, para. 2.

The second factor that contributes to the constant error in an examination mark is the difficulty of the questions set. In examinations which have to be prepared with the degree of difficulty suitable for the group examined[1], the examination should vary in difficulty roughly between questions which only 5% of the students are able to answer and questions which only 5% are unable to answer. According to Ruch, "The proper degree of difficulty can be defined as that of allowing the average (mean) score to be approximately half of the maximum possible score. No pupil should earn either a perfect of a zero score." [2] Again, Guilford defines the optimal difficulty of a test as that stage where the test is passed with a probability of 0.5 by the average candidate.[3]

In other examinations, however, rigid standards have to be observed in setting the level of difficulty of the paper. This is the case in college examinations and in such general public examinations as the Qualifying, Entrance, Matriculation, etc... Here, the degree of difficulty of an examination item should not be interpreted with reference to the number of candidates passing it in any particular year.

Methods applicable to the analysis of degree of difficulty of examination items are discussed in chapters III and V. Often an analysis of the results of an examination - with special reference to separate items- throws light on the degree of difficulty of questions, which may be a guide in the setting of other examinations in the future.

Granting that the errors due to the examiner's standard and to the degree of difficulty can be determined, scores obtained by candidates on an examination may be weighted to reduce the element of error. For instance, unless there is very close agreement among a candidate's marks by different examinars, the weighted mean[4], with the reciprocal of each examiner's standard deviation as weight, gives a better approximation to his true mark. Similarly, in summing for the final marks in an examination, questions can be weighted for their difficulty (more difficult questions given a larger proportion of contribution to the total) - unless the questions are homogeneous in difficulty, in which case it has been found that unweighted scores show very high correlation with weighted scores[5]. But weighting questions with a wide range of difficulty is desirable, and it follows that, if the difficulty of the various items on an examination range over the whole scale at the base of the normal curve [6], the marks on separate questions must be weighted before they add up to the totals.

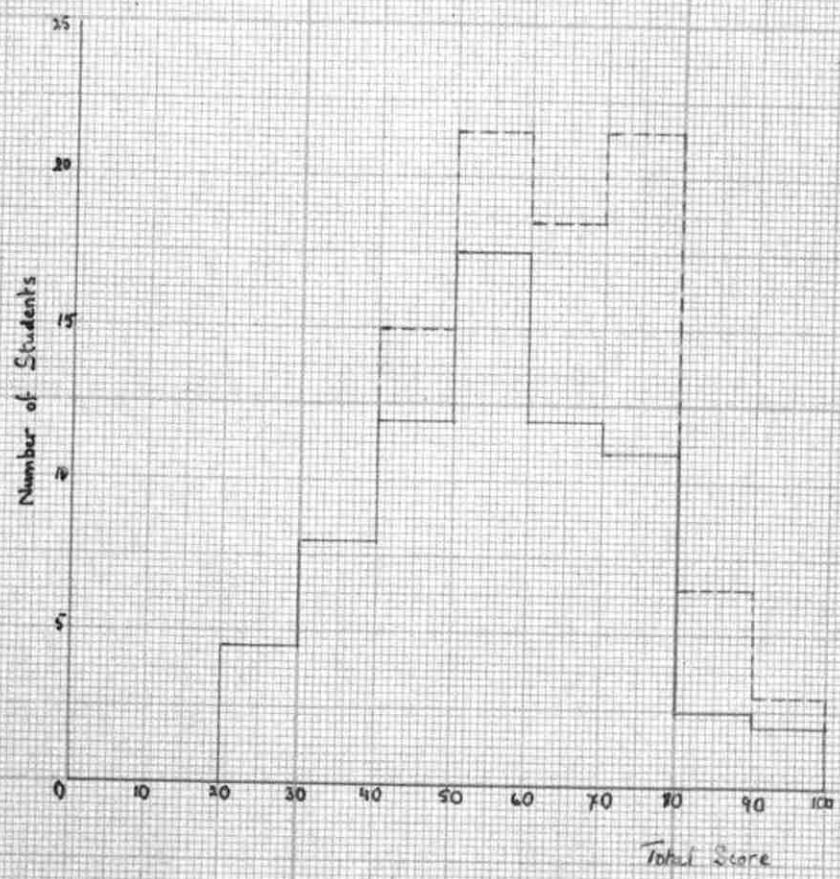
One method of evaluation that disposes of the constant error at the outset is the method of ranking. It has many supporters, including

---

(1) This is usually the case in ordinary school examinations in lower classes.
(2) The Improvement of the Written Examination, p. 121.
(3) Psychometric Methods, p. 445.
(4) for definition, see Ch. III B below.
(5) Monroe, De Voss & Kelley, p. 62; Spence, Chapter IV.
(6) Ch. III C, section 1.

Mathematics Quiz – Freshman, R.V.B. (1946)

Distribution of Scores
in the whole class and in lower 5 divisions.



- - - - - Whole class.
——— Divisions I–V only.

Ruch[1] and Vernon[2], and Spence has adopted it in his proposed scheme for improving college marking systems[3]. In ranking it is possible that some variable errors too are partially eliminated, those arising out of the necessity (in marking) of often assigning arbitrarily the evaluation of a paper on a minutely graduated scale of 100.

Although ranking is useful in schools for purposes of competition, marking is indispensable for examinations where the level of achievement is constant regardless of the performance of examinees. Besides, a disadvantage in ranking is the question of interpreting unequal increases in the intervals between the ranks, for the difference in performance value (the quantity measured by the mark) between the papers of the first and second candidates may exceed the differences between those of the ninth and tenth or others, whereas the rank difference is 1 in both cases.

Still another method used is that of grading, or assigning to the examination script a grade A, B, C, ..., thus partly evading the problem that arises out of accuracy of scale. But since each category or grade comprises usually a minimum of 10 marks on the % scale, the evaluation is only a rough one - and it may be best applied to supplement ordinary marking in borderline cases.

Marking out of a maximum of 10 involves the same problems as marking out of a maximum of 100, with the additional complication in having to deal with fractions!

Whichever method of evaluation is used, it appears after grouping the marks into frequencies that the distribution of abilities in an ordinary class of a fairly large size may be described by the normal "error" curve as an accepted approximation[4].

Such a distribution is, of course, expected when the group is a random selection as far as the ability examined is concerned. In specially selected groups the distribution is apt to present obvious skewness due to the elimination of a section of that group around one end of the curve in the distribution - the elimination being the result of selection. This will be found to be the case in a class after an entrance examination or in a division of a class after placement tests. The effect of selected groups on a distribution curve is illustrated in the accompanying graph.

If a group of candidates is examined for a particular job and part of the group selected as fit, this part will not, in general conform to the normal distribution if tested before long by a similar examination.

The application of the normal curve is desirable, even in slightly "abnormal" cases because of its symmetry of shape and facility in computation, the tables available for it reducing considerably the labour involved in other calculations of distribution.

---

(1) The Improvement of the Written Examination, p. 37.
(2) The Measurement of Abilities, Chapter II.
(3) cf. p. 12.
(4) evidence in Garrett, p. 89; Gregory, pp. 206-210; Rugg, pp. 207-230; Spence, Ch. IV.

For all practical purposes, only that section of the normal curve that lies between -2.5σ and $+2.5\sigma$ is used. This arbitrary "trimming" is allowed since it neglects only 1.24% of the measures[1]. The range is sometimes taken from -5σ to $+5\sigma$, and for use of this curve in biological experiments the zero point (or origin) is taken at -5σ; $5+z$ being taken as the "probit value" of the variable.

Although the normal curve cannot be safely used to describe the distribution of an examination results in a small class, it is sometimes applicable to a grouping of class sections, or more than one class given the same material by the same teacher, or given similar examinations. In a class, for instance, which is made up of several divisions, neither can the lower divisions nor the upper ones be expected to have means near the mean of the whole class taken together. The dispersions are, of course, smaller in every one of the divisions[2]. In such cases it may be appropriate to apply a normal distribution curve in order to supply the missing scores in any one of the divisions.

In a random sample of individuals, a trait especially well described by the normal distribution is intelligence. In a Civil Service Intelligence test given to 40000 candidates, Burt found it "definitely demonstrated" that intelligence was distributed in a way conforming very closely with the normal curve[3]. Wood has shown that intelligence scores correlate highly with achievement scores in school - a relation which is very helpful in revising distributions of assigned marks. This assumes, of course, that every student uses his natural abilities to the same degree, which unfortunately is not the case. However, intelligence scores may be used for guidance or to decide whether a child, if promoted, possesses the potential abilities to carry him to the required class level.

In fact, intelligence is a factor that enters into all work in school and, in particular, affects - to a greater or lesser degree - performance in all examinations, and it may be possible, given sufficient data, to isolate this factor and determine its weight in an examination by the usual process of factor analysis[4].

Vernon has shown how it is possible to analyse the whole of educational measurement into measures of a set of definite independent factors, which account for all test intercorrelations. This method, however, is best applied to standardised tests. His pattern of the analysis of tests into factors may be represented as -

Test	General Factor	Group Factors	Specific Factors
1	x	x	x
2	x	x	x
3	x	x	x
4	x	x	x
5	x	x	x
etc...	.	....	,

(1) -3σ to $+3\sigma$ neglects 0.27% of the cases, -4σ to $+4\sigma$ 0.0064% of them.
(2) cf. Ch. III C, section 5.
(3) Essays on Examinations, p. 101.
(4) Ch. III C, section 6.

the general factor being the element running through all performances of the set; group factors are common in some but not all of the tests, and these he detected after eliminating the general factor by the residual correlations among some subjects which tended to fall into correlated groups. In an analysis on the usual school subjects group factors appeared in:-

a) arithmetic problems and mechanical,
b) handwork, drawing, writing, speed, quality,
c) dictation, reading, speed,
d) composition, history, nature study, etc..

In addition, every subject possessed specific factors which were peculiar to itself and affected no other subjects.

The predictive value of a factor for any subject (i.e. the degree to which it can predict subsequent performance in that subject) can be measured by the correlation between scores awarded on a standardised test [1] in that factor and corresponding examination marks in the specified subject.

R.A.Kent and Esther Schuers of N. Western University carried some research on the "Predictive Value of Four Specified Factors for Freshman English and Mathematics"[2]. The factors were -

(1) Mental alertness score,
(2) High School quarter grades,
(3) Number of units of the subjects offered, and
(4) General achievement in all Freshman studies.

The results were not very encouraging and, in general, predictive value was not very high. But each factor of the above had some value for at least a part of the range of grades given.

Although such predictive measurements may prove to be of special value in the study of individual cases, their effect on examination standardisation does not repay the labour involved in isolating the factors and measuring their respective values in prediction, - and other more straightforward statistical methods are better applied and easier to follow.

---

(1) Obviously, some errors involved in the marking of the examination render this measurement less reliable than it should be; but since r is calculated with regard to deviations from the mean (Ch. III C, section 5), constant error arising in the central tendency does not affect its size, and with the grouping of the marks into classes with a range of 10 marks each a large proportion of the variable errors counterbalance.
(2) School and Society, Vol. 27, p. 242. (1928).

Chapter III

STATISTICAL METHODS.

A. Introduction

This chapter is a bare exposition of methods of statistics which may be used in standardising the setting, marking and distributing results of examinations. The statistical terms, statistics and parameters used are introduced in section B and defined, and the symbols used in later work indicated. In section C, the practical methods are listed, means of computation explained and illustrated - the illustrations being taken in most cases from analyses of fresh material from mathematics examinations. Since it was not possible to obtain data for illustrating all the methods surveyed, illustrations are given where they are necessary to elucidate the methods and where data was available, while some methods were only explained and hints given regarding possibility of application.

The presentation of every method is usually followed by a discussion of its assets and drawbacks. This discussion takes up the mathematical side of the problem - ignoring purely educational views, so that educational discussion is, as far as possible, eliminated, and only introduced where educational factors have special bearing on the methods surveyed and their application.

Standardising methods fall naturally into five main groups with regard to their specific functions:-

1. Those methods applicable to the preparation of the examination, the setting of the paper and determining difficulty of questions. In these methods the examiner usually relies on results and data collected from previous examinations similar in form or content.

2. Standardising by assigning expected types or scales of distribution results, used for establishing a common point of view between examiners (if there are more than one) or serving as flexible guiding marks for each examiner by himself.

3. Standardising scoring methods. These are numerous - all aiming to reduce the constant and variable errors of marking. Those surveyed in this section are only of a statistical nature, and typical of many others which are merely variations on them.

4. If scoring in not sufficiently standardised, scores are converted by any one of the methods next surveyed into marks that lend themselves to ready comparison [1].

5. After the papers are marked [2], a study of the distribution of results can reveal various characteristics of the examination, of the class examined, or of the marking - and show the kind and degree of error in marking. These will suggest ways for future improvement.

---

(1) This, according to Spence, is the best application of statistical methods for standardising examinations.
(2) The marking being made reliable by statistical or other methods.

B. Statistics and Parameters Used

Measures of Central Tendency.

m, the arithmetic mean, is the central measure in a set, and is the average of the measurements, defined mathematically as

$$m = \sum_i x_i / n, \quad (i=1,2,\ldots\ldots n) \text{ n being the total frequency.}^{(1)} \tag{1}$$

In a frequency distribution,

$$m = \sum_i f_i y_i / \sum_i f_i , \tag{2}$$

where f_i is the frequency in class i, y_i is the value of the centre of class i.

Similarly, the weighted mean is defined as

$$m = \frac{\sum_i w_i x_i}{\sum_i w_i} , \tag{3}$$

where w_i is the weight assigned to x

The mean of the parent population is similarly defined but is symbolised by μ.[2]

med, the median, is that value of the variable which divides the total number of variables (or total frequency), when arranged in ascending or descending order of value, into two equal parts. Mathematically,

$$\text{if } \sum_{i=1}^{k} f_i = \sum_{i=k}^{n} f_i = n/2 ,$$

then x_k is the median of the distribution.

The mode is the value of the variable with the maximum frequency.

---

(1) This holds also for the following summations.
(2) in most cases, where statistics are denoted by Latin letters, parameters are symbolised in Greek letters.

Measures of Scatter or Dispersion

The simplest measure of dispersion is the range which is defined as the size of the interval between the lowest and highest values.

i.e. range = l-a,
where l is the highest value,
a the lowest value in the distribution.

s^2, the variance of a sample distribution, is the arithmetic mean of the squared deviations $(x_i - m)^2$ from the mean m,

i.e. $$s^2 = \sum_i (x_i - m)^2 / n. \qquad (4)$$

It is also termed the "second moment" m_2 about the mean as origin.[1]

s, the square root of the variance, is the standard deviation, the most commonly used measure of dispersion

$$s = \sqrt{\frac{\sum (x_i - m)^2}{n}} \; . \qquad (5)$$

The variance and standard deviation of the population are symbolised by σ^2 and σ, respectively.

The coefficient of variation is used to judge the significance of a standard deviation with reference to the scale of measurement - since s does not have an absolute scale value but one depending on the units of measurement and on the range of measures. It is defined by

$$s/m \quad \text{or} \quad \sigma/\mu \quad \text{(for sample or population)}, \qquad (6)$$

and in percentage as $100s/m$ or $100\sigma/\mu$.[1]

The upper and lower quartiles, q_3 and q_1 are those values of the variate that divide the total frequency into $\frac{1}{4}$ and $\frac{3}{4}$ groups when arranged in ascending or descending order of magnitude,

i.e. if $\sum_{i=1}^{h} f_i = \sum_{i=l}^{n} f_i = n/4$,
then $x_h = q_1$,
$x_l = q_3$.

If analogously defined, q_2 is the median of the distribution.
q, the quartile deviation (or semi-interquartile range) is defined as

$$q = \frac{q_3 - q_1}{2} \; . \qquad (7)$$

The interquartile range $(q_3 - q_1)$ includes the middle half of the

---

(1) The sum of squared deviations from the mean m is (contd. p. 27)

variates, and therefore in any distribution the probability of a variate's assuming a value inside this range is $\frac{1}{2}$ - i.e. it has an equal probability of lying outside or inside the range. For this reason the q of the population was called the <u>probable error</u>, P.E.[3]

<u>Deciles</u> and <u>percentiles</u> are values in a distribution located in a manner analogous to quartiles. The deciles divide the variables into tenths, the percentiles into hundredths. Thus the 25th percentile is identical with q_1; the 50th percentile with the median, q_2 and the 5th decile; and the 75th percentile identical with the q_3 of the same distribution.

Measures of Asymmetry and Kurtosis

The standard measure of skewness (or asymmetry) of the distribution is in terms of the third moment m_3 (or μ_3) about the mean,

$$m_3 = \sum_i (x_i - m)^3 / n.$$

β_1, the measure of <u>skewness</u>, is then defined as

$$\beta_1 = m_3^2/m_2^3 \quad [4] \quad (\text{or } \mu_3^2/\mu_2^3 \text{ for the population}), \tag{8}$$

being positive or negative according to the sense of the skewness.

Pearson's measure of skewness depends on the fact that the mean and mode do not coincide in a skew distribution, and it is defined by

$$\frac{m - \text{mode}}{s} \quad [5] \quad . \tag{9}$$

Just as the third moment about the mean measures skewness, the fourth moment m_4 measures the degree of flattenning or excess at the centre of the distribution,

$$m_4 = \frac{\sum_i (x_i - m)^4}{n} \quad ,$$

---

(1)(contd.)
$$\sum_i (x_i - m)^2 = \sum x_i^2 + nm^2 - 2\sum x_i m$$
$$= \sum x_i^2 + \frac{(\sum x_i)^2}{n} - 2\frac{(\sum x_i)^2}{n}$$
$$= \sum x_i^2 - \frac{(\sum x_i)^2}{n} \quad . \tag{4a}$$

(2) Weatherburn, p. 7.
<u>N.B.</u> Formulas given without reference or derivation are found in any book on mathematical statistics - in particular those mentioned in the bibliography.

(3) In a normal distribution, P.E. = 0.6745 .

(4) The third moment is divided by s to standardise the scale.

(5) Dividing by s standardises the scale and gives an absolute measure.

and β_2, the measure of kurtosis, is defined as

$$\beta_2 = m_4/m_2^2 \text{ (or, for the population, } \beta_2 = \frac{\mu_4}{\sigma^4}). \qquad (10)$$ [1]

In the normal curve, $\beta_2 = 3$,
and therefore excess is measured by $\beta_2 - 3$ (in a leptokurtic curve)
and flattenning by $3 - \beta_2$ (in a platykurtic curve).

Measures of Correlation.

A correlation table gives frequencies in classes distributed with respect to two variables x and y taken on the two rectangular axes.

r, Pearson's product moment coefficient of correlation in a bivariate distribution is defined as

$$r_{xy} = \frac{\sum_i x_i' y_i'/n}{s_x s_y}, \qquad (11a)$$ [2]

where the x's and y's are taken as deviations from their means;

or

$$r_{xy} = \frac{\frac{\sum_i x_i y_i}{n} - m_x m_y}{s_x s_y}, \qquad (11b)$$

The subscripts denoting variables x and y.

The magnitude of r may be taken to measure the degree to which the association between x and y approaches a linear functional relationship.[3]

The line that best fits, by the method of least squares, such a linear relationship is the regression line of y on x[4], whose equation is

$$y = \frac{\sum_i x_i' y_i'/n}{s_x^2} x. \qquad (12)$$

The gradient of this line is b_{yx}, the coefficient of regression of y on x, and

$$b = \frac{\sum_i x_i' y_i'/n}{s_x^2}. \qquad (13)$$

The values of y corresponding to x on the regression line are estimates of the observed values on the scatter diagram. The sum of squares of deviations of these values from the estimated ones,

$$\sum_i (y_i - b_{x_i})^2,$$

---

(1) Aitken, p. 38.
(2) The numerator is the "covariance" of x and y.
(3) Weatherburn, p. 73.
(4) The equation of regression of x on y can also be determined in an analogous manner.

which is a minimum for the calculated b, can be divided by the number of these deviations, n, to give the variance of residuals. Its square root, the standard error of estimate, s_{est}, is defined as

$$s_{est} = s_y\sqrt{1-r_{xy}^2}\ .^{(1)} \tag{14}$$

The ratio of the standard error of the estimate of a variate to the standard deviation of that variate is the coefficient of alienation, k, which is

$$k = \sqrt{1-r_{xy}^2}\ .^{(2)} \tag{15}$$

This is a measure of lack of correlation, just as r is a measure of correlation.[3]

In trivariate distributions we can calculate the correlation between two variates (x and y, say) after eliminating the effect of the third variate z upon both of them. The coefficient of this partial correlation $r_{xy\cdot z}$ is defined as

$$r_{xy.z} = \frac{r_{xy} - r_{xz}\, r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}\ .^{(4)} \tag{16}$$

The correlation ratio, η_{yx}, is the measure of curvilinear regression of y on x and is calculated by passing a curve through the means of y arrays in the correlation table - corresponding to the fitting of a straight line by least squares in linear regression[5].

$$\eta_{yx} = \frac{1-\sum_i\sum_j(y_{ij} - \bar{y}_i)^2/n}{s_y^2}\ , \tag{17}$$

where y_{ij} is the value of y in the ith row and jth column, $\bar{y}_i$ is the mean of the ith column.

This is a measure of the degree to which the association between the variables approaches a single-valued functional relationship[6].

ρ, Spearman's coefficient of rank correlation, measures correlation in a bivariate sample, where the number of individuals is small, from the magnitude of the difference (d) between the ranks in the two sets. For n variables which assume different values in x and y series,

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}, \tag{18}$$

where d is the difference between the ranks of the same individuals under the x and y classifications.

The formula follows directly from the product moment coefficient formula.[7]

---

(1) Weatherburn, p. 72; Aitken, p. 89; ....
(2) Aitken uses the term "residual dispersion", p. 96.
(3) Ezekiel, p. 376.
(4) Aitken, p. 113; Weatherburn, p. 25; Kendall, p. 372.
(5) since the sum of squares of deviations from the mean is a minimum.
(6) Weatherburn, p. 90.
(7) Baten, p. 173; Kendall, p. 389.

The values of all coefficients of correlation range from +1 for perfect correlation to -1 for inverse correlation, 0 denoting absence of any correlation[1]. The significance of any observed value of r for a specified degree of freedom can be read from tables (Fisher & Yates, Table VI).

Sampling Errors.

The standard error, S.E., is defined as the standard deviation of residuals when the value of a variate is estimated from another[2]. If the mean μ of a population is estimated from the mean m of a sample, the standard deviation of m with the true mean μ as origin is the S.E. of the mean.
It can be proved[3] that the mean of a normal sample is normally distributed and that the standard error of the mean,

$$S.E._m = s/\sqrt{n} \quad , \tag{19}$$

where s is the standard deviation of sample variates,
n the number of variates in the sample.

Again, it can be shown that[4]

$$S.E._s = s/\sqrt{2n} \quad , \text{ with the same notation.} \tag{20}$$

Assuming a normal population, the significance of any error of statistics in a large sample can be expressed in units of its standard error, regarding the standardised statistic as significant if greater than 2.5 S.E.[5]

For small samples, and where normal distribution of statistics cannot be assumed, significance is tested by other special distribution.s.

Special Distributions

The Chi-square distribution is characterised by the probability density

$$dP = \frac{1}{\Gamma(\frac{1}{2}n)} (\tfrac{1}{2}\chi^2)^{\frac{1}{2}(n-2)} \exp(-\tfrac{1}{2}\chi^2) d(\tfrac{1}{2}\chi^2) \text{ ,}^{(6)}$$

or, for ν degrees of freedom,[7]

---

(1) Kendall, p. 393; Aitken, p. 87; Weatherburn, p. 73.
(2) cf. formula (14).
(3) Aitken, p. 128; Weatherburn, p. 110; Kendall, p. 224; Baten, p. 269.
(4) " ; Weatherburn, p. 137; " ; " .
(5) p. 21 above.
(6) Weatherburn, pp. 164-168.
(7) or number of independent (unrestricted) variates.

$$dP = \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} (\chi^2)^{\frac{1}{2}(\nu-2)} \exp(-\tfrac{1}{2}\chi^2) d(\tfrac{\chi^2}{2}), \qquad (21)$$

and $\chi^2 = \sum_i x_i^2/e_i$, (22)

where x_i's are deviations of n independent variates from their expected values,
e_i is the expected value of x_i .

With $\nu = 1$, the distribution reduces to

$$dP = \frac{1}{\sqrt{2\pi}} \chi^{-1} e^{-\frac{1}{2}\chi^2} \chi \, d\chi \quad (1)$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\chi^2} d\chi \text{ - a normal distribution.}$$

Since it is the sum of squares of deviations, χ^2 seldom or never assumes a value of zero. Tables prepared by Fisher (Fisher and Yates, IV) give for any number of degrees of freedom, ν , the probability that a certain value of χ^2 is exceeded in independent distributions. If the result is highly improbable, then the hypothesis of independence is discarded.

χ^2 is used as a <u>test of goodness of fit</u> of observed (O) to theoretical (T) values.(2) It is defined as

$$\chi^2 = \sum_i \frac{(O_i - T_i)^2}{T_i} \quad \text{as in (22) .} \qquad (22a)$$

If χ^2 is significant, the hypothesis of a non-fit is disproved. If the probability of its being exceeded in a non-fitting (independent) set is high, χ^2 is insignificant, and the fit is poor.

In measuring the significance of means for large samples, we use the fact that the ratio $(m-\mu)/SE_m$ is normally distributed. In small samples SE_m is not close enough in value to $\sigma/\sqrt{n}$, and the distribution of "<u>student's ratio</u>",

$$t = \frac{m-\mu}{SE_m} \quad \text{is symmetrical, leptokurtic, and} \qquad (23)$$

different for each n.(3) as n increases it approaches normality. Tables for t are available (Fisher & Yates, III) for any degree of freedom, giving the probability of t being numerically greater than the computed value.

Similarly, Fisher derived the <u>z-distribution</u> for small samples to test significance of the ratio of variances. For two samples whose variances (calculated for degrees of freedom) are s_1^2 and s_2^2 ,

---

(1) since $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
(2) $\sum \frac{(O-T)^2}{T}$ is distributed like χ^2 in large samples.
(3) Peters and Van Voorhis, p. 171; Aitken, p. 133.

$$z = \frac{1}{2} \log_e s_1^2/s_2^2 = \log_e s_1/s_2 .^{(1)} \qquad (24)$$

The distribution is independent of the standard deviation of the population, and for large samples approaches normality. Fisher's z tables (Fisher & Yates, V) show values of z that are exceeded in random sampling with probabilities of .2, .05, and .01, corresponding to specified values of ν_1 and ν_2 (degrees of freedom).

From these, Snedecor produced tables of F[2], the ratio of variances s_1^2/s_2^2, taking the ratio of the larger estimate of variance to the smaller - at the 5% and 1% levels.

The distribution of F is independent of the standard deviation of the population, its curve asymptotic to the X-axis on the right, and lies entirely on the positive side.

---

(1) Tippett, p. 117.
(2) "F" in honour of Fisher.

C. A Survey of Methods.

1. Setting: Difficulty of Questions

Diagnostic Curves

These can be constructed like Paterson's curves[1] for separate questions to determine their diagnostic values. In a large examination, considering general ability as adequately described by the candidate's total marks on the whole examination, the diagnostic curve for any particular question shows the regression of the performance of the class in that question on their general ability in the examination.

Class performance may be described either as:-

(a) the proportion of candidates passing[2] the question to those attempting it, or
(b) the mean or median of the candidates' marks on the question.

A diagnostic "curve"[3] may then be constructed as follows:-

1. Divide the class into groups according to their totals; into grades A, B, ... or intervals 0-9, 10-19, 20-29, ...
2. For every one of the groups in the previous step, calculate (a) or (b) above.
3. With total mark as abscissa and class performance as ordinate plot the values calculated in '2'.
4. Join the plotted points, and the resulting smoothed graph is the diagnostic curve.

Following Guilford,[4] we can consider the median of the curve in (a) as the level of difficulty of the question[5], and the slope as proportional to its diagnostic value, which implies that a good diagnostic question has a curve with a large slope, and a poor one does not show an appreciable rise for the higher grades.

In (b) the regression of the means can be similarly interpreted as an indication of the diagnostic value of the question.

---

(1) p. 10 above.
(2) "passing a question" is not rigidly defined, but it is here taken as obtaining over half the maximum mark allotted to the question.
(3) In practice, we may also plot a histogram or frequency polygon.
(4) Psychometric Methods, p. 430.
(5) measured in percentage units, it is the mean total mark of those students who pass the question with a probability of .5.

Illustrations[1]

(a) The diagnostic curves[2] for the marks of the first quiz show that

Question 1 does not differentiate highly between good and poor students, but it shows a definite positive correlation between proportion of higher scores and general ability.

Question 2 is easy, dealt with successfully by students with as low a total as 25-30[3].

Question 3 is difficult, since it is attempted by few, and those who succeeded have high totals. It is also a good diagnostic question.

Question 4 does not fail many students, and is therefore easy.

(b) The regression of question means on totals for the mid-year examination show that

Question 1 is valid but with rather a low diagnostic value.

Question 2 has a stable mean score for upper and lower levels but rises sharply in the middle and is therefore a good test of ability.

The big slope near the median of the curve for Question 3 shows a good diagnostic effect on that level.

Question 6 discriminates well between good and poorer students.

Question 7 is not a good diagnostic question because of the uncertain regression of means on totals, but it may be effective in sifting out the weakest section of the class.

Diagnostic curves are useful in the compilation of valid type questions, each with a known diagnostic value and level of difficulty, for use in standardising future examinations. But their construction requires additional work and their interpretations are based on the assumption of reliable marking. For this reason it is desirable to base them on results from large classes and after careful marking.

---

(1) First quiz and Mid-Year examination, Mathematics 101, A.U.B., 1946-7.
(2) Appendix B.
(3) Another factor affecting the results of these two questions is the fact that they appear early on the paper and therefore candidates are apt to give them more time. This may be remedied by changing the order of questions for random groups of the candidates.

Timing

Since very often a question on an examination is said to be more "difficult" than expected owing to the quantity of material it demands rather than any intrinsic complication, an evaluation [1] of its degree of difficulty can be made by an estimate of the time taken for answering it. This is especially the case when the question requires little more than memory or routine work. Then the difficulty value of the question may be taken as directly proportional to the time required for answering it [2].

This method has many limitations. In applying it to thought questions, a large error factor is introduced by assuming the examiner's mind to function along the same lines and with the same variations in rate for all questions as the examinee's. In any case, one examiner's estimate of time is too biased a measure for application in an objective study.

Illustration [3]

One examiner recorded the time taken by him to solve each of problems 1 to 6 of the examination, and obtained the following results:

Question	1	2	3	4	5	6
Time	9'	5'	3'	4.5'	6'	2.5'

Taking the means of students' marks on these questions as criteria of their difficulty, we can compare them with the examiner's timing. They were [4]-

Question	1	2	3	4	5	6
Mean mark	10.2	10.2	7.7	11.9	6.6	8.0

Ranking the questions in descending order of difficulty both according to time taken for solution and to mean results, we can calculate the coefficient of rank correlation.

Question	Rank by Time	Rank by Means	d
1	1	4	9
2	3	5	4
3	5	2	9
4	4	6	4
5	2	1	1
6	6	3	9

Then $\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 36}{6 \times 35}$

$= -1/35$, which shows even a slight negative correlation.

---

(1) besides subjective estimates by the examiner.
(2) for application of this method to mathematics, see Ch. V.
(3) Mid-Year examination in mathematics 101, A.U.B., 1946-7.
(4) for a maximum of 18 on each problem.

A remedy for this method may be the testing of questions on a number of people (of ability nearer to that of the students) and averaging the time taken by them.

Scaling of Difficulty by the Normal Curve

As in diagnostic curves, the proportion of students passing a certain question (being one way of describing the achievement of the class taken as a whole in that question) may be taken to indicate its difficulty level. In a fairly large class where one can safely assume a normal distribution of the ability examined, we can scale the difficulty of each question on the base of the normal curve as follows:[1]

1. Find the percentage p of students who fail each question.
2. On a normal curve, take the origin at the mean, and grade the base of the curve from -2.5σ to $+2.5\sigma$.
3. From normal distribution tables read the abscissa x corresponding to each value of p, where p represents the area under the curve between -2.5σ and x[2].
4. Convert x into percentage or any other desired scale. Then x is a measure of the difficulty of the question, deduced from the proportion of students failing it.

Illustration

For the same mid-year examination marks, with scale of difficulty 0 to 100, and mean difficulty at the origin 50,

$$\sigma = 20, \quad \mu = 50.$$

Question	Number of Failures	Number of Attempts	Proportion of Failures	Proportion - 0.5 [3]	x/σ	Difficulty Value.
1	70	173	.4046	-.0954	-.24	45.2
2	61	179	.3409	-.1591	-.41	41.8
3	53	98	.5408	.0408	.10	52.0
4	54	192	.2812	-.2188	-.58	38.4
5	85	110	.7727	.2727	.75	65.0
6	67	104	.6442	.1442	.37	57.4
7	60	78	.7692	.2692	.61	62.2

According to this, the ranking of the questions compared with their ranking according to the mean scores of separate questions, is

Question	1	2	3	4	5	6
Ranking by Means	4	5	2	6	1	3
Ranking by proportion of failures	4	5	3	6	1	2

$$\text{and } \rho = 1 - \frac{6 \times 2}{6 \times 35} = \underline{.943.}$$

---

(1) Method adapted from one used by Rugg on a standardised Algebra test.
(2) If the ogive of the normal curve is used, ordinates are read as p and abscissae as x.
(3) Correcting for origin.

After determining the degree of difficulty of a question, the examiner can use the information in many ways:-

1) He can select all questions on a paper of equal difficulty (or almost equal), if there is a choice of questions.

2) He can group questions of equal difficulty into sections and require at least one answer from each section.

3) If questions with highly different degrees of difficulty are set together on the same paper, the marking of each question must be weighted (in direct proportion to its difficulty) when the separate marks are added for the totals.

4) If questions are weighted for difficulty the student must be warned of this procedure - possibly by an indication of the maximum mark allotted to each question.

5) If difficulty has been determined by proportion of successes (or failures) the information can be kept to help in the choice of the problems or of similar problems for future examinations.

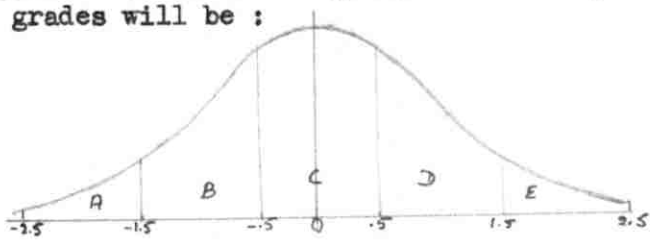
2. Expected Distribution of Results

Normal Distribution

If the expected distribution of results in a class is normal, the examiner knows beforehand the percentage of students that theoretically should receive any assigned mark or grade - the measure being the scale at the base of the normal curve, and the percentage corresponding to any range of values being 100 x portion of the area between the ordinates at both ends of the range.

For example, if the measure is in grades A, B, C, D, and E, one way of fixing the scale is by taking -2.5σ to 2.5σ, with the deviation of 1σ for each grade - as in the diagram.[1] Then the percentages in the respective grades will be :

A	6 %
B	25 %
C	38 %
D	25 %
E	6 %.



Or, in marking out of 100, we can fix the scale with 50 as the mean grade and ranging from 0 at -2.5σ to 100 at $+2.5\sigma$, then the percentage of students in the respective 10 marks intervals will be :

Class Intervals		% of cases
0- 9	90-99	2
10-19	80-89	4
20-29	70-79	10
30-39	60-69	15
40-49	50-59	19

The values were read from Vernon's curves which show proportion of measures which deviate from the mean in a normal distribution by any multiple of the standard deviation (read to 3 decimal places, and approximated in the above calculations to the nearest integers).

The use of such a guide in marking helps some examiners whose marking otherwise involves too much error - correcting for their standards and errors in dispersion. But, owing to the arbitrary boundary lines and the fact that this particular distribution is not necessarily always typical of the class it is applied to, examiners should not adhere too rigidly to the percentages calculated according to it, but rather take them as hints and as checks on their marking.

---

(1) Another example is a scale used in the Palestine Matriculation (see Ch. IV).

Distribution of Abilities in a Special Subject

The general achievement of any class in a particular subject may be arrived at by giving the class at the beginning of the year a standardised test in that subject, if such a test is available. The frequency distribution of marks obtained on that test describes the distribution of the particular ability in the class. The curve obtained can be used by the examiner to predict the results of a future examination on the same class in the same subject.

While this method does not fall into the error of arbitrarily assigning a theoretical curve of distribution, as the normal curve, it should not be followed too strictly and slight discrepancies in the second curve from the expected one should be accepted, since account has to be taken of other factors such as the reliability of the standardised test itself. Or, assuming high reliability of the test, we must also allow for a nonuniform change in the abilities of students during the lapse of time between the test and the examination, and for the fact that performance in an examination is not determined by ability alone but by preparation also. It is therefore important to regard the type distribution as merely a rough indication of the general form to be expected.

Distribution of Intelligence Scores

In a small class or a class where distribution of intelligence or general abilities is not necessarily normal, we can still make use of the distribution of intelligence scores obtained by the class. Knowing this distribution we can expect a similar one for achievement scores (with an adjacent mean and similar dispersion and skewness), when the two scores have high correlation in general.[1] In fact, the degree to which the marking guided by intelligence scores is reliable depends on the value of this correlation in the particular case to which it is applied. For, if we estimate achievement scores from intelligence scores, the standard error of estimate,

$$SE = s\sqrt{1-r^2}\ ^{(2)},$$

where s is the standard deviation of achievement marks.

For perfect correlation, $r = 1$, $SE = 0$,
for no correlation, $r = o$, $SE = s$, and the estimate of each score may be as much in error as the deviation of any other score in the distribution.

---

(1) p. 21 above.
(2) formula (14) above.

Marking Scales

These are also preassigned distributions, giving the percentage of students that should fall into each of the five grades A B C D E in the marking. Various forms of these scales are used, some based on a theoretical expectation (such as the normal distribution scale above) and others on the results of a large number of typical examinations.

The Dearborn Scale[1],

Grade	A	B	C	D	E
Percentage of Students	2	23	50	23	2

is approximately normal - being, in fact, the binomial distribution for n=9 and a probability of $\frac{1}{2}$.

The g.f. for a binomial distribution with $p=q=\frac{1}{2}$, n=9, is

$$G(t) = (\tfrac{1}{2})^9 (1+t)^9$$
$$= (1/496)(1 + 9t + 9x8t^2/2 + 9x8x7t^3/2x3 + \ldots\ldots t^9) .$$

Taking coefficients of t^0, t^1 to fall in class A,
" " t^2, t^3 " " " " B,
" " t^4, t^5 " " " " C,
" " t^6, t^7 " " " " D,
and " " t^8, t^9 " " " " E,

the distribution will be, for a population of 100,

Class	A	B	C	D	E
Number of Cases	1.95	23.4	49.3	23.4	1.95

which, to the nearest integer[2], is Dearborn's scale.

The Foster scale[1], based on a study of distribution of grades in several colleges, is

Grade	A	B	C	D	E
Percentage of Students	6	24	50	18	2

The curve of this distribution is slightly skew - which may be expected in college classes.

The Missouri plan[1] is given as

A	B	C	D	E
3	22	50	22	3

---

(1) Given in Spence, p. 7.
(2) in class C the 49.3 is altered to 50 in order to add up to 100 in the total frequency.

This is symmetrical with mean at C. Considering the grades to include each an equivalent range of marks, we can calculate the standard deviation of the distribution as follows:-

Grade	x	nf	nfx	nfx^2
A	2	3	6	12
B	1	22	22	22
C	0	50	0	0
D	-1	22	-22	22
E	-2	3	- 6	12
		100		68

$\Sigma fx^2/\Sigma f = 0.68$

therefore $\sigma^2 = .68 - .08$ (Sheppard's correction)

$= .6$

and $\sigma = .775$.

Fitting a normal distribution[1], we obtain frequencies of 3, 23, 48, 23, 3 in the respective grades, and the Missouri plan is therefore approximately normal.

The same arguments for and against using a normal distribution hold here - for the scales are arbitrary and of a theoretical value only, and an examiner must not apply them blindly in all cases. In some examinations (such as in large classes, with non-specialised material, etc...) the examiner may believe the results to fall according to one specific scale - or he may even make up a scale - and arrange his marking accordingly. There are examinations, also, in which only a specified percentage of students has to be passed, or a specified number of distinctions given. This means the fixing of a marking scale before marking.

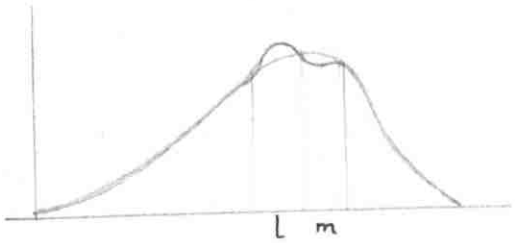
---

(1) for method of fitting see section 5 below.

3. Scoring Methods

Type Correction Plan[1]

This plan requires each institution to keep in record types of distribution curves for abilities in all its subjects. After an examination is corrected the distribution curve of the results is drawn up and "corrected" to fit the type curve it most approximately resembles, the corrections being done by remarking any doubtful papers where an excess of frequencies occurs in the grade - possibly arriving at a few papers that should, by the figure, belong to class m and were incorrectly given an l grade.



The idea behind this method is that the standards of instruction in any particular institution tend to remain more or less constant over a number of years.

Again, this method involves unnecessary arbitration, and the "corrections" may not be valid, especially in small classes. It is also doubtful whether any corrections done are worth the labour taken in doing them.

Scoring Guided by Previous Records.

The prognostic value of a previous record on a present examination is measured by the coefficient of correlation between that record and the corresponding present examination.

For the Freshman class at the A.U.B. (1946-7) the correlation in mathematics marks between the Placement Test and the Mid-Year Examination was 0.49, which shows a little prediction though not an outstanding one. In this case, we must take into account the fact that both examinations were liable to scoring errors. For this reason, we may expect better prediction if the "previous record" is derived from a standardised test.

r is calculated by eliminating the means from the two sets of marks, and therefore does not compare general level. The means must then be compared separately .

---

(1) Method proposed by Weld - given in Spence.

The cumulative index[1] of a class is calculated by

1. Calculating the average of each student's term[2] marks. This is the individual cumulative.
2. Taking the average of the class averages in "1" - the class cumulative.
3. Fixing a standard mark as unit for the whole institution.
4. Expressing every class cumulative in index form with reference to the standard in '3'. This is the cumulative index of the class.

In marking an examination for that class the teacher looks up the last cumulative index and allows his average mark to deviate only slightly from that index - except under unusual circumstances.

This allows the teacher to mark in accordance with the ability and usual performance of the class, and corrects for different levels in standards of the different teachers of the same class. But, since in the cumulatives, high and low marks in different subjects contribute equally to the totals, a class weakness or strength in any particular subject is not recognised and the corrections in the average later may involve errors in some subjects.

Gaps in Marking

This is a method applicable to passing or failing in borderline cases, advised to its examiners by the Palestine Matriculation Examinations Committee[3]. The examiners are at first asked to mark the scripts and then grade them (according to the marks) for distinctions (A & B), credit (C) and failing (below D). The minimum marks and the maximum marks for each grade are suggested by the committee, but the examiner is allowed reasonable deviations from these and is asked to set his own limit points, guided by any gaps in the consecutive marks near the limits suggested by the Committee.

"After the minimum marks for B, C, and D have been fixed, the examiner should read again all scripts within about 5 marks above and 5 marks below those minima, to see if any scripts within those limits deserve pushing up or down a grade ..... Examiners should then revise all scripts which are at the minimum or maximum of a grade, if there are no breaks in the consecutive marks, and add or deduct a few marks in order not to have so-called 'hard' cases ."

---

(1) Leker. cf. p. 13.
(2) or semester, or year.
(3) Instructions to Examiners, Section A 'Marking', July 1945.

Marking True-False Questions

An erro r committed in giving credit to all true answers in a True-False examination is that in most cases answers are made by guesswork - and since there are two alternatives the student gets the correct answer with a probability of 0.5.

To eliminate this it is sometimes suggested that the examiner gives credit corresponding to $\frac{r-w}{n}$, where r is number of correct answers, w is number of wrong answers, and n is total number of answers.

If s stands for the fraction of maximum scores in this method,

$$s = \frac{r-w}{n} = \frac{r-(n-r)}{n} = \frac{2r-n}{n} .$$

If the student answers all questions by guessing, the probability $= \frac{1}{2}$ for obtaining correct answers, and $E(r) = n/2$;

then $$E(s) = \frac{2(n/2) - n}{n} = 0 ,$$ and he gets no credit.

If he employs no guessing and knows the answers,

$$E(r) = n ,$$

then $$E(s) = \frac{2n-n}{n-1} = n/n = 1,$$ and he gets full credit.

Although this is only a linear transformation of the number of correct answers, it takes account of the unavoidable guessing factor by deducting marks for wrong guesses, and also by warning the students, which is an important psychological influence.

Soderquist's method of weighting scores according to the degree of assurance with which the student makes his responses is another method applicable to True-False items on an examination. On such a paper the questions are preceded by instructions like:-

> "Mark a statement + if you judge it to be true and 0 if false. You may claim credit up to 4 points for each of your responses if you wish. Before your response encircle 4, 3, or 2 depending on the credit you want. If your answer is wrong the penalty will be double the amount of credit you claim. If you claim no special credit, you should, nevertheless, answer all the questions. No-credit questions will be scored in the right minus wrong way."

The examiner then scores
Total credit claimed for correct answers - 2(credit claimed for wrong answers).

This takes care of eliminating guessed answers; but, on the other hand, it encourages guessing and wastes the time of the student in making him decide how sure he is about each question and answer.

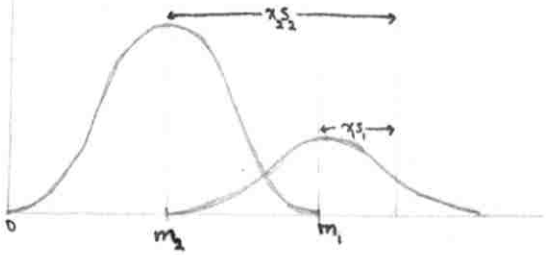
4. Conversion of Scores.

Linear Transformation

This method is used when the examiner can readjust his set of scores in level or scatter to a required preassigned standard.

Let the assigned distribution have mean at m_2 and standard deviation s_2, while the scores have mean at m_1 and a standard deviation of s_1.

Then any score can be expressed in standardised units with respect to either distribution - as x_2 and x_1,



then $\quad x_2 s_2 + m_2 = x_1 s_1 + m_1$,

and $\quad x_2 = x_1 s_1 / s_2 + \dfrac{m_1 - m_2}{s_2} \qquad$ I

This is the equation of a straight line with slope s_1/s_2, and with intercept $\dfrac{m_1 - m_2}{s_2}$,

and every score obtained can therefore be transformed to a mark on the standard distribution, if the constants of the two distributions are known[1].

When only the maximum and passing mark of the two distributions are known, a graphic device always facilitates the conversion. This may be constructed as follows:[2]

1. Set up a frequency distribution of scores obtained. Calculate m and s.
2. Locate the lowest passing score (at -1.5 s, for example).
3. With the lowest passing score at the origin, lay off in equal steps on the horizontal axis of a graph the range of possible scores.
4. With the lowest passing mark in the standard system at the origin, lay off on the vertical axis of the graph distances corresponding to the possible marks.
5. Locate on the graph the point corresponding to the maximum possible mark and maximum possible score.
6. Draw a straight line through this point and the origin. Then, to find the mark corresponding to a particular score, read up to the diagonal line and the point where it cuts it has for its ordinate the required mark.

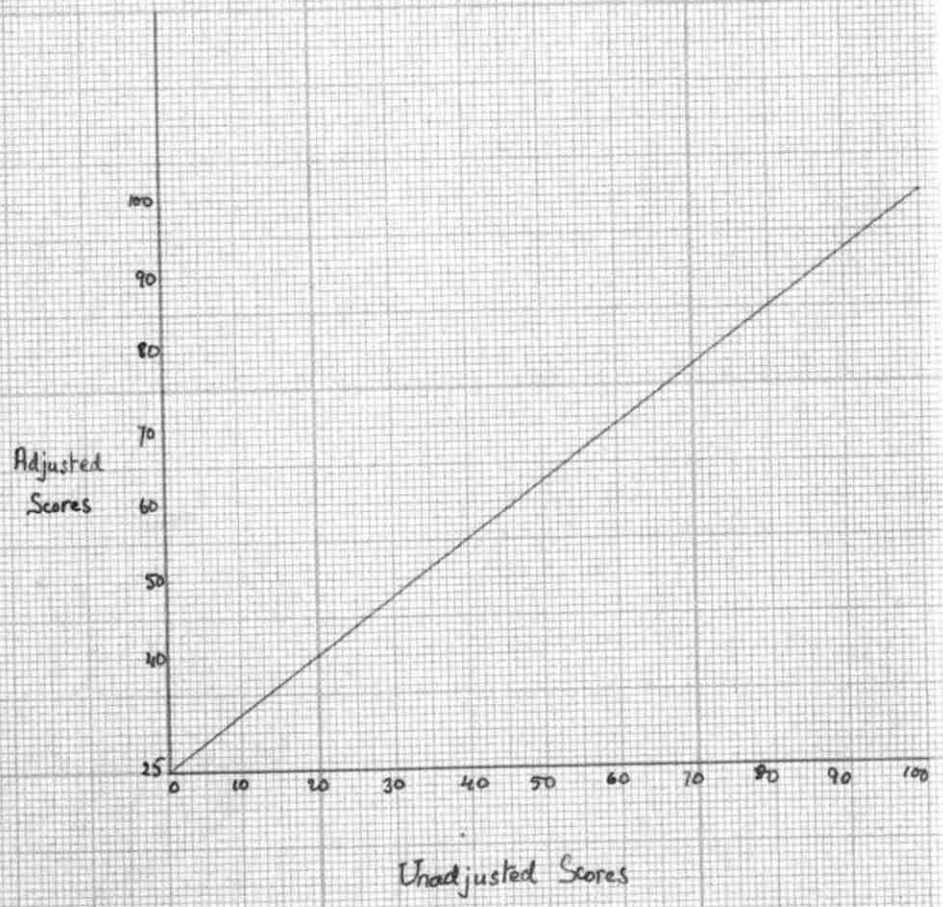
---

(1) In practice, examinations are not marked with scores in standardised units.

(2) Method followed by Sims to reduce constant errors.

LINEAR TRANSFORMATION

(illustration, p.46)



Origin at lowest possible scores

Such a conversion corrects for some constant errors of measurement by correcting for level and dispersion - provided that the standard distribution is the one more appropriate for the group examined.

Illustration[1]

The scores obtained on the mid-year examination were rather low and every mark was readjusted by use of the transformation -

adding to the score $\frac{1}{4}$ the difference between it and 100,

that is, adjusted score = original score + (100 - original score)

or $$a = u + 25 - u/4$$
$$= \tfrac{3}{4}u + 25,$$ which is a linear transformation.

Comparing it with equation I, p. 45, the intercept of the line is 25, and the slope $\frac{3}{4}$ or 0.75.

The two frequency distributions were -

Class Interval	x	f_u	f_a	xf_u	x^2f_u	xf_a	x^2f_a
0- 9	-5	1	0	-5	25	0	0
10-19	-4	10	0	-40	160	0	0
20-29	-3	20	0	-60	180	0	0
30-39	-2	35	11	-70	140	-22	44
40-49	-1	48	33	-48	48	-33	33
50-59	0	39	52	0	0	0	0
60-69	1	25	57	25	25	57	57
70-79	2	9	28	18	36	56	112
80-89	3	5	9	15	45	27	81
90-99	4	0	2	0	0	8	32
		192	192	-165	659	93	359

Then $m_u = 45.8$ $m_a = 59.8$
$s_u = 16.1$ $s_a = 13.4$.

$$m_a/s_a - m_u/s_u = 4.5 - 2.9 = 1.6,$$

or, in units of u, $1.6 \times 16.1 = 25.8$.

The actual change in origin is 25; the difference is probably due to inaccuracy in calculating the constants from class groupings of frequency.

(1) Mathematics 101, A.U.B. , 1946-7.

Transformation to a Normal Distribution

This is done by the use of a converting scale similar to the one proposed by Stevasón[1]. Since we deal with percentage scores ranging from 0 to 100 we construct the scale with a similar range over -2.5σ to $+2.5\sigma$. Each point on the graduated scale denotes the percentage of the total area below the curve to the left of the ordinate erected at that point. Also, the points q_1, q_3 and $\pm\sigma$ are located. After construction, the scale is pasted on cardboard and cut along the ruled edges.

For converting the scores, we can use one of two methods suggested by Stevason.

1. The Quartile Method:

On a ruled paper draw up a frequency distribution, adjusting spaces between the lines so that the whole range fits the marking scale in total lenghth.

Calculate q_1 and q_3 and draw parallel lines through these points across the page. Adjust the marking scale to the angle necessary to make its quartile points fall on the q-lines of the table. With the marking scale in this position read the marks corresponding horizontally to any score.

2. The Standard Deviation Method:

In the same way, after finding the mean and standard deviation of the distribution, locate the two points $m\pm s$ on the table.

Adjust the placing of the marking scale so that the points $\mu\pm\sigma$ fall on the horizontal lines through m+s and m-s, and then read off corresponding marks as before.

According to Stevason, this method is better suited for skew distributions, because it gives a wider spread of marks than the quartile method.

Percentiles

A percentile graph can be used to convert scores to percentile ranks[2]. The graph is constructed by use of the frequency distribution of scores. For every class boundary (of scores) the percentage from the total frequency is calculated of those who fall in the distribution below the limit of the boundary. On a graph with abscissae as percentiles, and ordinates as scores, points corresponding to every class limit are plotted - and the points joined by a smooth curve which is the percentile graph. From this the percentile rank that corresponds to any particular score in the range can be readily determined.

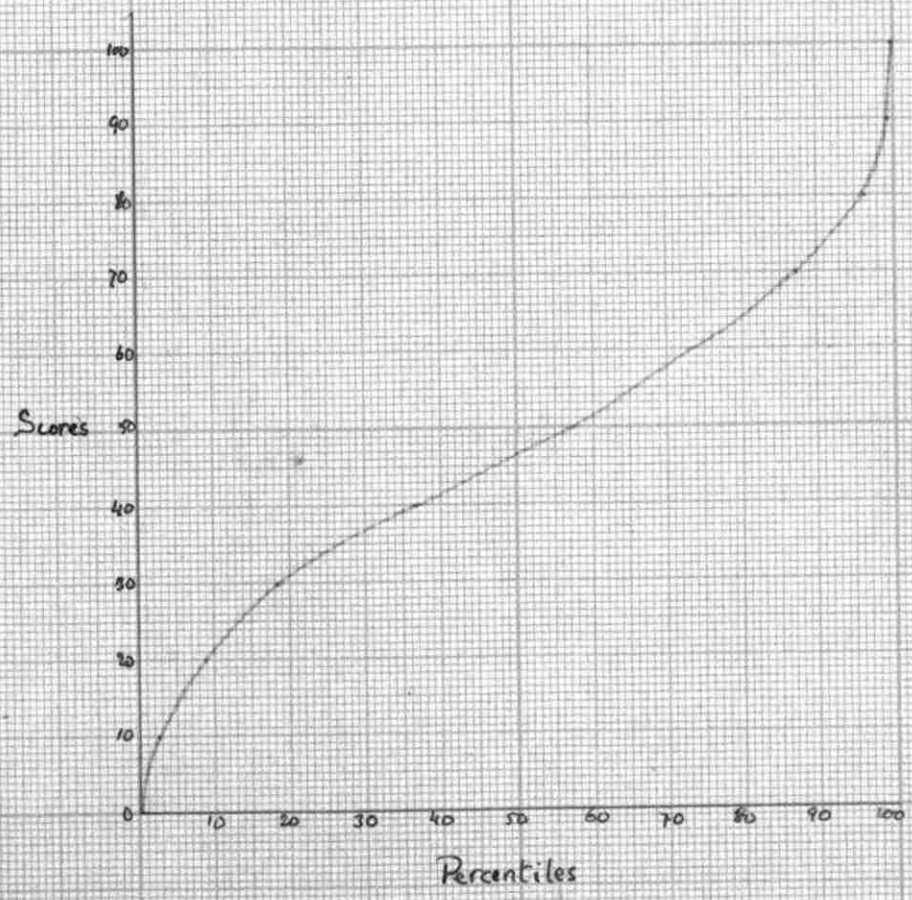
---

(1) p. 11 above.

(2) note (1) p. 48.

Percentile Graph

(illustration – p. 48)



Illustration[3]

Class	Upper Limit	Frequency	Cumulative F.	Cumulative F. as % of total
1	10	10	10	2.7
2	20	23	33	8.9
3	30	34	67	18.1
4	40	69	136	36.7
5	50	77	213	57.6
6.	60	57	270	73.0
7	70	54	324	87.5
8	80	33	357	96.5
9	90	12	369	99.7
10	100	1	370	100

The graph may also be used for reading any mark corresponding to a particular decile or percentile.

If the converted distribution is taken as standard, then any other distribution can be made to fit to it by its deciles. After calculating the deciles of the new distribution, a reading of the graph gives the "standard" scores for those decile ranks, or for any intermediate percentile.

The advantage of this method of conversion over Stevason's scale is that the assumed standard is not necessarily normally distributed, but has any distribution decided as being the best applicable.

T-Scale

This is a scale for marking proposed by McCall[2] based on the transformation of ranks or scores to a normal distribution. By weighting raw scores according to the frequency of students obtaining any particular grade of scores, it provides a basis for a uniform marking scheme in an institution. The method is, briefly, -

1. The frequency of students in every class interval of scores is counted.
2. Considering the fact that the number of students with better scores affects the value of that score, the number

$$b = f/2 + F ,$$

where f stands for frequency in any interval,
F stands for the sum of frequencies in higher intervals,

is computed for every class interval.

(1) Sorenson pp. 105-115; Garrett, p. 46; Vernon, Ch. IV.
(2) Spence - The Improvement of College Marking Systems.
(3) Palestine Matriculation Mathematics Examination - 1945.

3. The percentage 100 - 100b/n is found.

Or, if the results are given in ranks, b is the given rank, and the percentage in '3' is calculated from it.

Assuming a normal distribution of scholastic ability, and taking the normal distribution curve from -5σ to $+5\sigma$ (with origin at -5σ), the percentage in '3' is regarded as that section of the area under the curve which is bounded on the right by the ordinate erected at the probit corresponding to b. The probit value is then read from normal or probit tables. The T-score, which is a standard measure, is

100 x probit value of the rank.

This method, which standardises measures in an institution, provides for comparison within classes, between subjects, or between classes. But it assumes normal distribution in all groups, which is not a safe assumption in all these cases.

Difficulty Ratio[1]

This involves the weighting technique in expressing the average for any particular individual in all subjects as a ratio which shows the difficulty of his obtaining high grades - and which makes comparison easier between students who take subjects with different levels of difficulty. To obtain the ratio,

1) express each individual candidate's marks as grades in the A, B, C, ... scale,
2) calculate the frequency of students in each subject obtaining grades as high as A, B, C, ... ;
3) the reciprocal 1/f of this frequency is taken as weight w for each grade in each subject;
4) the weighted average of each individual's marks is calculated;
5) the simple average of each individual's marks is calculated.

Then the difficulty ratio is defined as:-

$$DR = \frac{\text{weighted average}}{\text{unweighted average}} \times 100$$

and is less than 100 when a student's programme is easier than average, greater than 100 when a student's programme is harder than average.

---

(1) Introduced by Spence.

Illustration

Let student A's marks in subjects M N O P Q be 78, 46, 80, 94, 35, respectively.

Taking the grades scale,

A	90-100
B	80- 89
C	70- 79
D	60- 69
E	40- 59
F	0- 39,

and denoting the frequency of grades as high as C in subject O by ${}_{C}f_{O}$, let the respective frequencies be

$$
\begin{array}{cccccc}
{}_{A}f_{M} & {}_{B}f_{M} & {}_{C}f_{M} & {}_{D}f_{M} & {}_{E}f_{M} & {}_{F}f_{M} \\
{}_{A}f_{N} & {}_{B}f_{N} & {}_{C}f_{N} & {}_{D}f_{N} & {}_{E}f_{N} & {}_{F}f_{N} \\
{}_{A}f_{O} & {}_{B}f_{O} & {}_{C}f_{O} & {}_{D}f_{O} & {}_{E}f_{O} & {}_{F}f_{O} \\
{}_{A}f_{P} & {}_{B}f_{P} & {}_{C}f_{P} & {}_{D}f_{P} & {}_{E}f_{P} & {}_{F}f_{P} \\
{}_{A}f_{Q} & {}_{B}f_{Q} & {}_{C}f_{Q} & {}_{D}f_{Q} & {}_{E}f_{Q} & {}_{F}f_{Q}
\end{array}
\quad .
$$

Then A's weighted average will be

$$
WA = \frac{\dfrac{78}{{}_{C}f_{M}} + \dfrac{46}{{}_{E}f_{N}} + \dfrac{80}{{}_{B}f_{O}} + \dfrac{94}{{}_{A}f_{P}} + \dfrac{35}{{}_{F}f_{Q}}}{1/{}_{C}f_{M} + 1/{}_{E}f_{N} + 1/{}_{B}f_{O} + 1/{}_{A}f_{P} + 1/{}_{F}f_{Q}} \quad .
$$

The simple average is

$$SA = (78 + 46 + 80 + 94 + 35)/5$$

and $$DR = 100\ WA\ /SA\ .$$

This ratio may be useful in colleges where students of the same class take different subjects and their final grades are then considered together and compared.

5. Analytic Study of a Distribution

Central Tendency and Dispersion

The mean of a distribution indicates the general level of marking and the standard deviation its dispersion.

In calculating the standard deviation from frequencies grouped into classes, the frequency is taken to be concentrated at the central value of the class. If the frequency density in a class is uniform, such a calculation involves no error - but in general the half of the class nearer the mean contains the larger proportion of the frequency. A positive error is therefore involved in calculating the second moment by frequency concentrated at the centre of the class, and Sheppard's correction for this grouping in the second moment is [1]

$$- w^2/12 \quad ,$$

where w is the breadth of the class interval.

Illustration [2]

Taking class intervals of 5 units over the % range, we calculate the mean and second moment by the summation method [3]

X	nf	x	Σ	Σ^2	Σ^3
0-	1	0	195	[illegible]	
5-	4	1	194	1695	
10-	7	2	190	1501	8062
15-	12	3	183	1311	6561
20-	10	4	171	1128	5250
25-	12	5	161	957	4122
30-	10	6	149	796	3165
35-	21	7	139	647	2369
40-	14	8	118	508	1722
45-	13	9	104	390	1212
50-	28	10	91	286	821
55-	19	11	63	195	535
60-	10	12	44	132	340
65-	10	13	34	88	208
70-	11	14	24	54	120
75-	7	15	13	30	66
80-	1	16	6	17	36
85-	1	17	5	11	19
90-	2	18	4	6	8
95-	2	19	2	2	2
	r!		1	1	2

(1) Aitken, pp. 44-46.
(2) Scores on Freshman mathematics Placement Test, A.U.B., October 1946.

The sums at the heads of the columns are $nm'_{(r)}/r!$

Therefore $$m'_{(1)} = \frac{1695 \times 1}{195} = 8.7$$

$$m'_{(2)} = \frac{8062 \times 2}{195} = 82.7$$

Transforming to ordinary moments,

$$m'_1 = m'_{(1)} = 8.7$$
$$m'_2 = m'_{(2)} + m'_{(1)} = 91.4.$$

For X as grouped, the mean is $8.7 \times 5 + 2 = \underline{45.5}$

The second moment about the mean is

$$91.4 - (8.7) = 15.7$$

For X, $m_2 = 15.7 \times 25 = 392.5$

Applying Sheppard's correction,

$$m_2 = 392.5 - 25/12 = 390.4,$$

and $s = \underline{19.65}$ (approximately).

The coefficient of variation[4] is $s/m = \underline{.432.}$

The moments calculated for distributions on separate questions of a paper show the type of question and its difficulty.

<u>Illustration</u>[5]

For questions 1, 3, and 4 the results were[6] -

m	13.206	14.955	7.19
s	8.2	9.4	7.6

---

(3) The ordinary power method is illustrated in the example on p. 46. By the summation method we obtain factorial moments, which can be transformed (Aitken, p. 41 & 44) to ordinary moments by simple relations. The factorial moment $m_{(r)}$ is defined as $\Sigma(x-m)^{(r)}/n$ where $z^{(r)}$ denotes $z(z-1)(z-2)(z-3).....(z-r+1)$.

That the sum at the head of the rth column is equal to $nm'_{(r)}/r!$ is demonstrated in Aitken, Appendix 2.

(4) formula (6), section B.

(5) Placement Test, Freshman A.U.B. , 1946. (four questions)

(6) Every question was marked out of a maximum of 25.

The mean and standard deviation were calculated only for those students who attempted the question - disregarding the fraction of the class that did not attempt it. Taking the class as a whole, these figures stand for the general results if we give the mean score on a question for any student who did not attempt it. This involves an error by raising the mean of difficult questions that are attempted by fewer students.(1) To eliminate this, we can calculate the constants of the distribution, giving 0 on any unattempted question.(2)

Let M_1 and s_1 be mean and standard deviation of the distribution as calculated,

M_0 and s_0 the mean and standard deviation giving zeros on unattempted questions;

then for question 1,

$$M_0 = M_1 \times \frac{187}{195} = 13.206 \times \frac{187}{195} = 12.66.$$

$$s_0^2 = \frac{187}{185}(s_1^2 + M_1^2) - M_0^2 = 71.35.$$

For the other questions of the test, the alteration resulted in -

Question	Number of Blanks	M_1	M_0	s_1^2	s_0^2
1	8	13.2	12.66	67.3	71.35
3	6	14.9	14.5	88.5	92.4
4	49	7.2	5.38	58.2	53.4

---

(1) If, on a question, only f ~~can~~ of the candidates present any work at all, and the rest are given a credit of M_1 each, the difference in the means if those are given zeros will be $M_0 - M_1$,

But $M_0 = fM_1$,

Therefore $M_0 - M_1 = -M_1(1-f)$.

The second moment $s_0^2 = f(s_1^2 + M_1^2) - M_0^2$

$= fs_1^2 + fM_1^2 - f^2M_1^2$,

And $s_0^2 - s_1^2 = fs_1^2 + fM_1^2 - f^2M_1^2 - s_1^2$

$= -(1-f)s_1^2 + (1-f)fM_1^2$

$= (1-f)(fM_1^2 - s_1^2)$.

Therefore, the second moment, and consequently s, will

increase if fM_1^2 exceeds s_1^2, and
decrease if fM_1^2 is less than s_1^2,

if blanks are given zeros.

(2) This cannot be fully justified, particularly in the last question, because a student may leave it unattempted because of lack of time and not out of the difficulty of the question.

If a grouped distribution is approximately normal, an estimate of the standard deviation can be obtained by the formula

$$s = N_1^2/2N_2\sqrt{\pi}\ .$$ [1]

Illustration[2]

Class Interval	Frequency (f)	f^2
0-	1	1
5-	4	16
10-	7	49
15-	12	144
20-	10	100
25-	12	144
30-	10	100
35-	21	441
40-	14	196
45-	13	169
50-	28	784
55-	19	361
60-	10	100
65-	10	100
70-	11	121
75-	7	49
80-	1	1
85-	1	1
90-	2	4
95-	2	4
	$N_1 =$ 195	$N_2 =$ 2885

$N_1^2 = 38025$, $N_2 = 2885$, $\sqrt{\pi} = 1.773$,

$$s = \frac{38025}{2885 \times 1.773} \times \sqrt{5} \qquad \text{(interval } w = 5)$$

$$= \underline{16.3.}$$

As previously calculated (on page 52), $\underline{s = 19.65}$. The difference may be due to the abnormality of the distribution, which is platykurtic.

A second measure of dispersion, the range, is sometimes used in describing frequency distributions. But it is better applied in cases where the density is more or less uniform in the distribution, since it overestimates the dispersion in distributions with freak cases. Ruch, for instance, reproduces the distribution of marks given by 115 teachers to a geometry paper.[3] The range of marks is 92-28 = 64 - a very wide range in a scale of 100 . Actually, as may be seen by reference to the

---

(1) Appendix A, para 4. N_1 is the sum of class frequencies,
N_2 the sum of squares of class frequencies.
(2) Placement Test, Freshman A.U.B., October, 1946.
(3) see Ch. V below.

frequency histogram, the range of the majority of cases is 92-51 = 41.

Although the sum of squared deviations from the mean is affected considerably by the cases at 28 to 31, it still gives a better picture by taking due account also of intermediate cases, which vary a great deal in weight.

Frequency Graphs

A histogram of a grouped distribution is constructed from its frequency table by taking along the ~~two~~ X-axis equal intervals corresponding to the class intervals of the distribution - and erecting on each interval a column along the Y-axis proportional to the number of cases in it.

The frequency polygon gives a more continuous picture of the distribution and is constructed from the histogram by joining the middle points of the bars (tops of columns) which represent the frequencies in each class.

These give a graphic representation of the facts, which is often sufficient to describe peculiarities in the distribution and degree of departure from normality.

Illustration[1]

A study of the frequency distribution graphs of separate questions may show ~~chew~~ characteristics of the question like the following:

Question 1: The distribution shows a U-curve, which is expected in single questions which a student either can or cannot solve. The distribution is also symmetrical.

On question 3 four definite marks are given. This may be due
either to the division of the question into 4 parts,
or to the teacher's personal effect, or both.
Forty % of the students have a perfect score; it is therefore comparatively easy.

Question 4 shows too many low scores
either because of its difficulty,
or because it is the last question, and most students only started at it.
Those who could tackle the question got high scores, and we find no in-between cases; it has a decisive dividing mark between pass and fail.

---

(1) Placement Test, Freshman A.U.B., 1946.
Histograms and copy of question paper in Appendix B.

The <u>sloped frequency curve</u>[1] combines in construction a histogram and a frequency polygon. It is constructed from a histogram by giving each bar at the end of any frequency rectangle the slope of the line joining the centres of the two adjacent bars - since, in general, there is a marked tendency for items to cluster toward the mode. This is accompanied by a slight decrease in dispersion[2].

An accurate study of the abnormalities of the curve - in particular, skewness and kurtosis - is made by the use of higher moments[3]. Sign and degree of skewness is measured by β_1, while β_2 gives any excess in the central portion.

<u>Illustration</u>[4]

class breadth w = 5, assumed mean at 42.

X	x	nf	Σ	Σ^2	Σ^3	Σ^4	Σ^5
0-	-8	1	1	1	1	1	1
5-	-7	4	5	6	7	8	9
10-	-6	7	12	18	25	33	42
15-	-5	12	24	42	67	100	142
20-	-4	10	34	76	143	243	385
25-	-3	12	46	122	265	508	893
30-	-2	10	56	178	443	951	(1368.5)
35-	-1	21	77	255	(570.5)		
40-	0	14	(84) (111)				
45-	1	13	104	390	(1016)		
50-	2	28	91	286	821	2155	(4157.5)
55-	3	19	63	195	535	1334	3080
60-	4	10	44	132	340	799	1746
65-	5	10	34	88	208	459	947
70-	6	11	24	54	120	251	488
75-	7	7	13	30	66	131	237
80-	8	1	6	17	36	65	106
85-	9	1	5	11	19	29	41
90-	10	2	4	6	8	10	12
95-	11	2	2	2	2	2	2
		195	0!	1!	2!	3!	4!

(1) Davies and Crowder, p. 79.
Illustrated in Appendix B.
(2) compare with Sheppard's corrections.
(3) formulae 8 & 10 above.
(4) Placement Test, Freshman A.U.B., 1946.
Calculations by central factorial moments (Aitken p. 43)

Central factorial moments are :-

$$mm'_{\{1\}} = 390-255 = 135$$
$$mm'_{\{2\}} = (570.5 + 1016) \times 2 = 3173$$
$$mm'_{\{3\}} = (2155 - 951) \times 6 = 6624$$
$$mm'_{\{4\}} = (1368.5 + 4157.5)24 = 132624.$$

Transforming to ordinary moments, (for x) ,

$$m'_1 = 135/195 = 0.6923$$
$$m'_2 = 3173/195 = 16.27$$
$$m'_3 = (6624 + 135)/195 = 34.66$$
$$m'_4 = (132624 + 3173)/195 = 693.32.$$

Transforming to moments about the mean (for x),

$$m_1 = 0 \qquad m'_1 = 0.6923$$
$$m_2 = 15.794$$
$$m_3 = 2.6368$$
$$m_4 = 646.09 .$$

For X (as grouped),

Mean at $42 + 0.692 \times 5 = 45.46$

$$m_2 = 15.794 \times 25 = 394.8$$
$$m_3 = 2.637 \times 125 = 329.6$$
$$m_4 = 646.09 \times 625 = 403806.25.$$

Applying Sheppard's corrections,

Mean at 45.46

$$m_2 = 394.85 - 25/12 = 392.77$$
$$m_3 = m_3'' = 329.60$$
$$m_4 = 403806.25 - 25/2\, m''_2 + \frac{7 \times 625}{240} = 390888.86.$$

$$\beta_1 = m_3^2/m_2^3$$
$$= (329.6)^2/(392.77)^3$$
$$= \underline{.00179} , \qquad \text{or } 1/571 .$$

$$\beta_2 = m_4/m_2^2$$
$$= 398888.86/(392.77)^2$$
$$= \underline{2.585.}$$

Therefore the distribution is platykurtic.

A rough measure of skewness can be obtained from Pearson's measure[1]

$$\text{Skewness} = \frac{m - \text{mode}}{s}$$

by use of the approximate relation between mean, mode and median[2] which may be symbolised as

$$\text{mode} - m = 3(\text{med} - m)$$
$$\text{mode} = m - 3(m - \text{med}).$$

Substituting in the above expression,

$$\text{Skewness} = \frac{m - (m - 3(m - \text{med}))}{s}$$
$$= 3(m - \text{med})/s.$$

Illustration

In the above example,

$$\text{mean} = 45.46,$$
$$\text{med} = 42.5$$
$$s = \sqrt{392.77} = 19.8,$$

$\therefore$ Skewness $= 9/19.8 = 0.46$, as compared with .002 in accurate calculation.

A study of some frequency distributions may show the mean, modé and median to approach each other closely and the general shape of the curve to resemble that for a normal distribution. It is possible to test how far a normal curve can "fit" this distribution by first finding the normal curve of best approximation to it, where the relative class frequencies are represented by segments of area under the curve between ordinates corresponding to the given class boundaries. This is done by setting the mean μ of the normal curve at the arithmetic mean m of the group of marks, and taking its standard deviation σ equal to the standard deviation of the marks. Expressing class boundaries in standardised units as deviations from the mean, the section of the area between any ordinate (class boundary) and the Y-axis can be read from the probability integral tables. Differences between consecutive readings give the sections of area corresponding to class frequencies. For a class of n students, n x the area section gives number of students who should fall into that class by the normal distribution.

---

(1) formula (9), section B.
(2) "In many probability curves of slight or moderate skewness the median lies between the mode and the arithmetic mean, nearly twice as far from the mode as from the mean." - Aitken, p. 31.

Illustration[1]

Scores grouped into classes of 10.

The mean and standard deviation are found by the usual methods. As calculated,

$$m = 44.64,$$
$$s = 16.15.$$

x	$z=\frac{x-m}{s}$	$\frac{1}{2}$ erf z[2]	$\frac{1}{2}\Delta$ erf z	$n(\frac{1}{2}\Delta$ erf z)	Observed Frequency
0	-2.77	-.4972			
			.0134	3	1
10	-2.14	-.4838			
			.0481	9	10
20	-1.52	-.4357			
			.1145	22	20
30	- .92	-.3212			
			.1617	32	35
40	-. 04	-.1595			
			.2888	55	48
50	.33	.1293			
			.1723	35	39
60	.97	.3340			
			.1066	22	25
70	1.56	.4406			
			.0488	9	9
80	2.18	.4854			
			.0120	3	5
90	2.80	.4974			

The fifth column shows the normal frequency distribution with the same parameters. To test the goodness of fit, χ^2 is calculated for the differences between observed and theoretical frequencies.

In this case, $\nu = 7$[3].

$$\chi^2 = 1.333 + .182 + .262 + .89 + .445 + .408 + 1.333$$
$$= \underline{4.783.}$$

For $\nu = 7$, $\chi^2 = 4.783$, the probability of obtaining a value of as great or greater is 0.60 (app.), or 60 %.

Therefore it is a very good fit.

---

(1) The fitting of a normal curve to the unadjusted scores on the mid-year examination, Mathematics 101, A.U.B., 1946-7.

(2) $= \int_0^z e^{-\frac{1}{2}z^2}dz.$

(3) Out of the 10 degrees of freedom 3 were used to determine the fitted curve (m, s, & n).

Coefficient of Correlation

The correlation coefficient r is used in various applications, such as:-

a) Measure of the degree of prediction,[1]
b) Coefficient of reliability [2] (self-correlation),
c) An indication of variable errors of marking of different examiners (by intercorrelating their markings of the same papers),
d) A measure of relationship between two subjects.

Illustration[3]

Paper I (x)

Paper II (y)

	0- -4	10- -3	20- -2	30- -1	40- 0	50- 1	60- 2	70- 3	80- 4	90- 5	f_y	y	yf_y	y^2f_y	Σx	$y\Sigma x$
0- -4	6	3		1	1						11	-4	-44	176	-34	136
10- -3	5	8	7	2	3						25	-3	-75	225	-60	180
20- -2	1	11	16	8	7						43	-2	-86	172	-77	154
30- -1		10	15	16	8	7	1	1			58	-1	-58	58	-64	64
40- 0		1	6	16	25	10	6	1			65	0	0	0	- 6	0
50- 1		1	6	17	27	30	8	2			91	1	91	91	20	20
60- 2			2	4	11	13	28	14	2	1	75	2	150	300	116	232
70- 3					2	4	6	10	2	1	25	3	75	225	59	177
80- 4							2	6	4	3	15	4	60	240	53	212
90- 5									5		5	5	25	125	20	100
f	12	34	52	64	84	64	51	34	13	5	413		138	1612	27	1275
x	-4	-3	-2	-1	0	1	2	3	4	5						
xf_x	-48	-102	-104	-64	0	64	102	102	52	25	27					
x^2f_x	192	306	208	64	0	64	204	306	208	125	1677					
Σy	-41	-67	-58	-17	20	61	89	83	51	17	138					
$x\Sigma y$	164	201	116	17	0	61	178	249	204	85	1275					

$_xm_1' = 27/413 = .0654$ $\qquad$ $_ym_1' = 138/413 = .3341$

$_xm_2 = 1677/413 - (.0654) = 4.06$ $\qquad$ $_ym_2 = 1612/413 - (.334) = 3.79$

$\therefore\ s_x = 2.015.$ $\qquad$ $\therefore\ s_y = 1.947.$

$$r_{xy} = \frac{\Sigma xy - m_x m_y}{s_x s_y}\ ^{(4)}$$

$$= \underline{.781}$$

(1) example on p. 20 above.
(2) Ch. II above.
(3) Palestine Matriculation Mathematics Examination (1942) - Papers I & II.
(4) Formula (11b).

Another method for calculating r is by the summing of frequencies along diagonals of a correlation table - obtaining the frequency distribution of (x-y). Then r is found after calculating the standard deviation of this distribution and of the two separate distributions by the relation [1]

$$r_{xy} = \frac{1}{2}\,\frac{s_x^2 + s_y^2 - s_{x-y}^2}{s_x\, s_y} \quad .$$

Illustration

In the above example we calculate the diagonal distribution x-y.

z=x-y	-4	-3	-2	-1	0	1	2	3	4	Σ
f_z	3	13	51	104	143	63	27	7	2	413
zf_z	-12	-39	-102	-104	0	63	54	21	8	-111
$z^2 f_z$	48	117	204	104	0	63	108	63	32	739

$$s_z^2 = 739/413 - (111/413)^2 = 1.717.$$

$$\therefore \quad r_{xy} = \frac{1}{2}\,\frac{(4.06 + 3.79 - 1.72)}{2.015 \times 1.947}$$

$$= \underline{.78} \quad .$$

If r is calculated from interval groupings (as is usually the case in its application to examination results), we can correct for the error thus introduced by the formula

$$r_{x'y'} = r_{xy}\,\frac{\sigma_{x'}\sigma_{y'}}{\sigma_x\sigma_y} \quad , \quad [2]$$

where r_{xy} is correlation in terms of class intervals,
$r_{x'y'}$ correlation in terms of variates,
& σ's are standard deviations in variates.

But it is seldom worth carrying through these calculations because the use of r itself here does not call for all this accuracy.

A narrow range of variation (in one or both variables) tends to reduce the coefficient of correlation.[3] Since correlation is determined by the position of paired values in relation to their means, r is positive or negative according to whether both values are on the same or opposite sides of their means. In a narrow range the probability is

(1) Appendix A, para 7.
(2) Appendix A, para 5.
(3) Sorenson, p. 272.

greater for paired values to fluctuate on either side of the mean than in a wider range - because change of position with respect to the mean is effected through a small fluctuation for most of the pairs in the distribution.

Again, since r is defined by the expression,

$$r_{xy} = \frac{\Sigma xy - m_x m_y}{s_x s_y},$$

a limitation of the range of x or y or both tends to decrease the covariance, which exceeds any other factor in the expression, and a diminution of the covariance diminishes r.

Illustration[1]

The correlation between Q and E (taking Q as x and E as y) was calculated by the usual method for the whole class and the result was

$$r = .486.$$

Correlation was again calculated between the same two sets of scores in isolated sections of the class and the correlations were -

in divisions I & II[2]	r = .4066,
in divisions V & VI	r = .3824,
in divisions III & VII	r = .4167.

In the three subdivisions r as calculated was less than the r calculated for the whole class. The range of marks in every case and its effect onrcan be read from the table below.

	Whole class	I & II	V & VI	III & VII
Range for x	20 - 100	40 - 100*	20 - 80	20 - 90
Range for y	20 - 100	30 - 100	20 - 90	20 - 90
Correlation Coefficient	.486	.4066	.3824	.4167.

* with the exception of 1 candidate in 20-29.

(1) Mathematics 101, A.U.B., 1946-7, first quiz (Q) and first term exercises (E) scores.
(2) The class was divided into 7 divisions in the beginning of the year after a placement test in which mathematics had a contribution of 1/3.

The probable error of r is given as

$$PE_r = .6745 \frac{1-r'^2}{\sqrt{n}} \text{ [1]},$$

where r' is the coefficient of correlation in the population.

According to Guilford, "the general rule is that r must be at least 4 times its PE in order to be significant. When r is very low, it should be more than 4 times its PE to be indicative of any correlation at all[2]."

But, since the value of r is obtained from the sample and it may be very different from the population coefficient r', a serious error is involved in substituting the former for the latter in the above formula. Besides, the use of PE in this case is not valid because the distribution of r' is not normal except for small or moderate values of the coefficient and for large n. As r' approaches .8, the sampling curve becomes very skew and the PE device is decidedly inapplicable[3].

Fisher's transformation of r to z'[4], where

$$z = \tanh^{-1} r = \tfrac{1}{2}(\log_e(1+r)-\log_e(1-r)) ,$$

can be read from tables[4] and then the standard error or PE device can be applied because z is approximately normally distributed, especially for large samples, with variance $1/\sqrt{n-3}$.

Illustration[5]

For divisions V & VI, the correlation coefficient was found to be $r = .3824$. $n = 56$.

From Fisher's tables, corresponding to this coefficient,

$$z' = .403,$$

$$SE_{z'} = 1/\sqrt{53} = .1374,$$

$z'/SE_{z'} = 2.933$, which is, therefore, significant.

For comparison, if we use SE ,

$$SE_r = .8538/\sqrt{56},$$

$r/SE_r = 3.533$. This shows a higher level of significance, which may be misleading in drawing conclusions.

A disadvantage of z' is that it is only an intermediate statistic with no direct meaning.

---

(1) Guilford, Psychometric Methods. (2) Ibid. p. 333.
(3) Tippett, p. 176
(4) VII, Fisher & Yates. Dashed z is used by Tippett to distinguish between it and z in formula (24)
(5) First Quiz and Exercises in Mathematics 101 - (cf. example above)

Another distribution of r makes use of the distribution of Student's ratio. Here,[1]

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \quad .$$

To test the significance of r, we assume the variates to be uncorrelated, and calculate t. A significant value of t then discredits this assumption.

Illustration:

Taking the data of the last example,

$$t = \frac{.3824}{\sqrt{.8538}}\sqrt{54} = 3.042.$$

$\nu = 56 - 2 = 54.$

From table III in Fisher and Yates, for 54 degrees of freedom, the probability of obtaining as great a value or greater is less than 1%, and the correlation is therefore significant.

The coefficient of partial correlation shows the correlation between two sets of marks under the influence of a common factor after eliminating the effect of that factor.

For x and y, keeping z constant,

$$r_{xy.z} = \frac{r_{xy} - r_{xz}\, r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} \quad .$$

Illustration[2]

To calculate the correlation between first (x) and second (y) quarter grades, after eliminating the effect of the placement test (z), we calculate at first the total correlations:-

$$r_{xy} = .8298$$
$$r_{xz} = .4875$$
$$r_{yz} = .4928 \; .$$

Then $r_{xy.z} = \dfrac{.8298 - (.4875)(.4928)}{\sqrt{(1-.4875^2)(1-.4928^2)}}$

$= \underline{.776}$.

---

(1) Weatherburn, p. 193.
(2) First and Second Quarter marks, Mathematics 101, A.U.B., 1946-7.

Regression Coefficient

This measure is similar to the correlation coefficient but is best confined to examples of the effect of one variable on the other. Numerically, b_{yx} is equal to the slope of the regression line of y on x.

Illustration[1]

The correlation coefficient calculated by the product moment method, was found to be .495. Taking x, the independent variable, to stand for Placement Test marks and y for the first Quiz marks, the regression, as calculated,

$$b_{xy} = r_{xy} s_x / s_y \qquad (r_{xy} = .495,\ s_x = 1.643,\ s_y = 2.005.)$$
$$= \frac{.495 \times 1.643}{2.005}$$
$$= \underline{.4056.}$$

b_{yx} is greater than r - because of the smaller dispersion of x.

Correlation Ratio

η^2 has more than one application - among them being :-

1) It shows the presence or absence of some law between two variates,
2) It shows the extent of its strength,
3) It tests the goodness of fit of any regression line[2].

In linear regression,

$$\sigma_{est}^2 = \sigma_y^2(1 - r_{xy}^2) \quad .$$

Assuming an equal standard deviation of all columns[3] in a correlation table,

$$\sigma_c^2 = \sigma_y^2(1 - \varepsilon_{yx}^2)$$

where σ_c^2 is variance of columns, ε^2 is unbiassed correlation ratio.

$$\therefore \quad \varepsilon_{yx}^2 = 1 - \frac{\sigma_c^2}{\sigma_y^2} \quad . \qquad \text{I}$$

For any column c,

$$\sigma_c^2 = \frac{\sum y_c^2}{n_c} - (\sum y_c / n_c)^2.$$

Summing and dividing by n (the total frequency),

$$\frac{\sum n_c \sigma_c^2}{n} = \frac{\sum(\sum y_c^2)}{n} - \frac{\sum n_c(\sum y_c/n_c)^2}{n}$$

or $$n\sigma_c^2/n = \sum\sum y_c^2/n - \sum n_c m_c^2/n$$

$$\therefore \quad \sigma_c^2 = \sigma_y^2 - \sigma_{m_c}^2$$

$$\varepsilon_{yx}^2 = \frac{\sigma_{m_c}^2}{\sigma_y^2} \quad \text{(by substitution in I)} \qquad \text{II}$$

---

(1) Placement Test and First Quiz, Mathematics 101, AUB, 1946-7.
(2) Method given in Peters & Van Voorhis, pp. 227-40
(3) Or "homoscedasticity".

Illustration[1]

		Divisions										
		I	II	III	IV	V	VI	VII	f_y	y	yf_y	y^2f_y
	20-						1		1	-3	-3	9
	30-		1			4	3	4	12	-2	-24	48
Scores	40-		1	5	3	3	10	7	29	-1	-29	29
on	50-	6	7	7	4	6	9	10	49	0	0	0
Quiz	60-	11	11	10	11	11	4	2	60	1	60	60
	70-	10	6	5	7	4	2	3	37	2	74	148
	80-		3		2				5	3	15	45
	f_c	27	29	27	27	28	29	26	193		93	339
	Σy_c	31	29	15	28	8	-11	-7	93			
	$\Sigma y_c^2/f_c$	36	29	8	29	2	4	2	110			

$$_ym_t^2 = (93/193) = .2322$$
$$_ym_a = (339/193) - .2322 = 1.5243,$$

$$\eta_{ya}^2 = \frac{110/193 - .2322}{1.5243} = .2216$$

$$\eta = \underline{.4707}.$$

An unbiassed correlation ratio is obtained from η^2 by the relation

$$\varepsilon^2 = \frac{(n-1)\eta^2 - (k-1)}{n-k}$$ [2],

where k is the number of columns,
n is the total frequency.

Peters & Van Voorhis give ε^2 tables which show the distribution of ε^2 when the true correlation is 0. The columns read for k-1 degrees of freedom and the rows for n-k degrees of freedom - and the table gives the maximum ε^2 that could be expected on the 5% and 1% levels on the basis of chance fluctuations when true correlation is zero.

Illustration

In the above example, unbiassed $\varepsilon^2 = \frac{192\eta^2 - (7-1)}{193 - 7} = \underline{.1965}$

---

(1) Analysis of Variation in marks on First Semester Quizzes with respect to class division, Mathematics 101, AUB, 1946-7.
(2) Appendix A, para 6.

Or, using formula I above,

$\varepsilon^2 = 1 - s_c^2/s_y^2$, where s_c is the standard deviation within a column, assuming homoscedasticity.

$= 1 - (S_y^2 - S_m^2)/S_y^2$, where S_m^2 is the sum of squares of the means of columns.

$$s_y^2 = 339/192 - (93/193)^2 = 1.5336$$

$$S_m^2 = f_c\Sigma(\Sigma\bar{y}_c^2/f_c)^2 - (93)^2/193$$
$$= 110 - 45$$
$$= 65$$

$$\therefore \quad S_c^2 = 1.54 \times 192 - 65$$
$$= 296 - 65 \qquad \text{(approximately)}$$

$$\text{and} \quad s_c^2 = \frac{296 - 65}{186}$$
$$= 1.236$$

$$\therefore \quad \varepsilon^2 = 1 - \frac{1.236}{1.5336}$$
$$= 1 - .8061 = \underline{.194}$$

$k-1 = 6$
$n-k = 186.$

At the 5% level = .032 to .043) smaller than .194;
At the 1% level = .052 to .069)

Conclusions:-

1) The result is highly significant;
2) Correlation is $\sqrt{.194} = \underline{.445}$

Rank Correlation

The coefficient of rank correlation ρ is applicable to small classes for cases where students are ranked. To calculate ρ between two sets of marks given to the same class,

1) rank each individual according to his standing in each distribution,
2) for each individual obtain d, the difference in his rankings in the two cases,
3) apply the formula for Spearman's coefficient,

$$= 1 - \frac{6\Sigma d^2}{n(n^2-1)} .$$

Illustration[1]

Student	Rank in P.T.	Rank in Q I	d	d^2
1	1	20	19	361
2	2	6	4	16
3	3	25	22	484
4	4	10	6	36
5	5	5	0	0
6	5	4	1	1
7	7	25	18	324
8	8	11	3	9
9	9	12	3	9
10	9	20	11	121
11	11	17	6	36
12	12	24	12	144
13	12	1	11	121
14	12	2	10	100
15	15	13	2	4
16	16	13	3	9
17	16	2	14	196
18	16	7	9	81
19	16	23	7	49
20	16	7	9	81
21	16	17	1	1
22	22	19	3	9
23	23	7	16	256
24	24	22	2	4
25	25	13	12	144
26	25	13	12	144
27	27	27	0	0
28	28	27	1	1
				2742

$$\rho = 1 - \frac{6 \times 2742}{28(784-1)}$$
$$= \underline{.25} \quad \text{(approximately).}$$

Pearson's transformation of ρ into r_ρ, where

$$r_\rho = 2 \sin (\pi\rho/6),$$

corrects for the spacing between ranks by changing it to normal spacing. This measure, however, is less reliable; it is, in fact, the least reliable of the three measures r, ρ and r_ρ, r being the most reliable[2].

The significance of ρ may be interpreted like that of r.

---

(1) Placement Test and First Quiz, Division I, Mathematics 101, A.U.B., 1946-7.
(2) Guilford, p. 341.

Analysis of Variance

The technique of Analysis of Variance [1] is that of dividing up a total sum of squared deviations of a variate from its sample mean into several distinct sums of squares each corresponding to a source, real or suspected, of variation.[2]

If the group of variates is classified into classes c, then, in symbols,

$$\sum_x (x - m_x)^2 = \sum_c f_c (m_c - m_x)^2 + \sum_c \sum_x (x - m_c)^2 \qquad \text{I}$$

where the left hand side is total sum of squares,
first term on RHS is sum of squared deviations of class means from the grand mean,
second term on RHS is sum of squared deviations of variates from their class mean.

Again, if the group is classified into classes c and d according to two different criteria of classification, then,

$$\sum_x (x - m_x)^2 = \sum_c f_c (m_c - m_x)^2 + \sum_d f_d (m_d - m_x)^2 + \sum_c \sum_d \sum_x (x - m_c - m_d)^2 \qquad \text{II}$$

On RHS, the first term is sum of squared deviations of c means from grand mean,
the second term is sum of squared deviations of d means from the grand mean,
the third term is the sum of squared deviations of variates from their means within classes - or "residual sum of squared deviations".

After calculating the sums of squares, and finding the degrees of freedom for each source of variation, the analysis of variance is set out in a table

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Square[3]
Total			
Between means of c's			
Between means of d's			
.....			
Residual			

(1) Introduced by Fisher in 1923.
(2) as defined in Aitken, p. 136.
(3) Mean Square = Sum of Squares/Degrees of Freedom (cf. Variance).

Snedecor's $F = \frac{\text{greater mean square}}{\text{smaller mean square}}$, and for a random distribution into classes F is not significant.

Fisher's $z = \frac{1}{2} \log F$ is given in tables V (Fisher & Yates) at the 20%, 5%, and 1% levels of significance.

Illustration[1]

The marks were classified by one criterion - the 7 questions. The distribution is as follows:-

Student	I	II	III	IV	V	VI	VII	
	Questions							
1	6		17	12	8		5	
2	12	11		12	8	4		
3	14			14	10	23	13	
4	6	8	16	18			7	
5	10	18		16	10		20	
6	18	16		12	17	23		
7	14	15		18	8		20	
8	12		13	12		4	5	
9	12	10		10	9		10	
10	12	9		14	8	12		
11.		8		16	8	2	6	
12	6	13		16		23	5	
13	12		16	14	16	6		
14	6	8	2	18		2		
15	14		13	16		0	5	
16	9	6		6		23	4	
17	10	10		18		2	0	
18	7	15		18	3	4		
19	16	7		2	7			
20	12	9	16	10	6			
21	9	13	12	18	7			
22	6	17		4	13	23		
23	12	10	12	18		8		
24	8	9	2	12	12			
25	8	17		14	18		5	
26	12	12		16	17		8	
27	12		16	11	8	8		
f_x =	26 +	21 +	11 +	27 +	19 +	16 +	14 = 134	Σf_x
Σx =	277 +	241 +	135 +	358 +	193 +	167 +	113 = 1484	$\Sigma(\Sigma x)$

(1) Mid Year Examination, Mathematics 101, AUB, Division 1, 1946-7.

Total sum of squares $= \Sigma x^2 - (\Sigma x)^2 / \Sigma f$ (1)
$= 20180 - 16435$
$= \underline{3745}$

Sum of squares between means of questions
$= \Sigma(\Sigma x_i)^2 / f_{x_i} - (\Sigma x)^2 / \Sigma f$
$= 16736 - 16435$
$= \underline{301}$

Sum of squares within questions
$= 3745 - 301$
$= \underline{3444}$

Analysis of Variance

	Sum of Squares	Degrees of Freedom	Mean Square
Total	3745	133	28.16
Between means of questions	301	6	50.17
Within questions	3444	127	27.12

$F = 50.17/27.12$
$= \underline{1.85}$

In F tables, 20 % point at 1.45
5 % point at 2.17
1 % point at 2.95

∴ the difference in mean squares is <u>not significant</u>.

(it is significant only at the 20 % level).

For comparison, the analysis of variance in Division IV of the same class gave --

	Sum of Squares	Degrees of Freedom	Mean Square
Total	5110	135	37.85
Between means	1252	6	208.67
Within questions	3858	129	29.86

$F = 208.67/29.86$, which is significant even at 1% level.

The reason for its significance in this division may be the fact that weaker students (in fourth division, as compared with those of the first) are more liable to be diagnosed by questions of different difficulty.

---

(1) Formula (4a) above.

Peters and Van Voorhis recommend ε^2 for educational research, since "constructive is just ready to begin where analysis of variance leaves off".[1]

In the illustration on page 66 above, an anlysis of variance reveals that -

$$s_c^2 = 1.236$$
$$s_m^2 = 10.833.$$

Then $F = 10.833/1.236 = 8.76.$

From F tables, the minimum F significant at the 5% level is 2.14 to 2.16,
at the 1% level 2.89 to 2.92,
which is much less than the observed F.

Therefore the variance between divisions is significant.

This is less information than that given by ε^2(p67).

If a variate z is the sum of several independent variates, x_i, then the variance of z is equal to the sum of the variances of the x's.

Let $z = x_1 + x_2 + x_3$, the variates being taken in standardised form. Squaring and summing, then dividing by n, the number of values each variate takes,

$$\begin{aligned}
\Sigma z^2/n &= \sum_{x_1}\sum_{x_2}\sum_{x_3}(x_1^2 + x_2^2 + x_3^2 + 2x_1x_2 + 2x_1x_3 + 2x_2x_3)/n \\
&= \frac{\Sigma x_1^2}{n} + \frac{\Sigma x_2^2}{n} + \frac{\Sigma x_3^2}{n} + \frac{2\Sigma\Sigma\Sigma(x_1x_2 + x_1x_3 + x_2x_3)}{n} \\
&= \Sigma x_1^2/n + \Sigma x_2^2/n + \Sigma x_3^2/n + \frac{2(r_{x_1x_2} + r_{x_1x_3} + r_{x_2x_3})}{n} \\
&= \Sigma x_1^2/n + \Sigma x_2^2/n + \Sigma x_3^2/n \qquad \text{(the r's are 0)}
\end{aligned}$$

$$\therefore \quad s_z^2 = s_{x_1}^2 + s_{x_2}^2 + s_{x_3}^2 ,$$

which can similarly be proved for any number of component variates, $x_1 \ldots\ldots\ldots x_\ell$.

This fact can be applied to test for the presence of any correlation between sets of marks for the same individuals in different subjects, or between different questions on an examination.

Illustration[2]

Variance of the total marks on the test was calculated and found to be

$$s_t^2 = 392.77.$$

For the separate questions, variances were -

---

(1) Statistical Methods, p. 358.
(2) Placement Test, Freshman Mathematics, A.U.B. , October 1946.

$$s_1^2 = 70.9$$
$$s_2^2 = 95.4$$
$$s_3^2 = 92.4$$
$$s_4^2 = \underline{53.4}$$

Therefore sum of
variances = 312.1, and $s_t^2 \neq \sum_i s_i^2$.
The hypothesis of independence between questions is therefore rejected.

$s_t^2 > \sum s_i^2$, which indicates positive correlation between the questions. This must be due to the correlation between abilities on the different problems.[1]

Control Charts

In drawing a control chart for a set (or several sets) of marks, the following is assumed :-

(a) that the frequency distribution curve of marks is monomodal, and similar to a normal curve;
(b) any definite deviation from the median or mean varies as $1/\sqrt{n}$ for any particular grouping,

for $s_{med} = 1.2533 \ldots\ldots s_m$ [2] (in samples from a normal parent)
$= 1.2533 \ldots\ldots s/\sqrt{n}$,
where s is the standard deviation of the sample,
n is the number of individuals in it.

Since, in a normal distribution, less than 0.3% of the cases fall outside the range $m \pm 3s$, less than 0.3% of the medians of samples should fall outside the range Med $\pm$ 3 s_{med} ,

where Med is the median of the samples combined,

and, if we draw the lines $y = \text{Med} \pm 3s_{med}$, they serve as control lines for the distribution - within which most of the medians should fall.

By assumption (b), the equations of the control lines are

$$y = \text{Med} \pm \frac{3 \times 1.2533\ s}{\sqrt{n}} \ .$$ [3]

(1) In fact, some of the correlations may be negative - although their sum is positive - due to competition between questions for time.
(2) Kendall, p. 225.
(3) Sandon uses, instead of $3s_{med}$ the range between the 1st and 9th deciles. Different dispersion measures are applicable to different examinations.

With abscissa $1/\sqrt{n}$ and ordinate y, the control lines are drawn. Then every subgroup i of the set (classified into subgroups by any number of criteria) is represented on the graph by the point ($1/\sqrt{n}$, med). Points falling outside the control lines show significant deviations.

An illustration of a control chart is given at the end of chapter IV (on Matriculation mathematics papers) with four criteria of classification.

To interpret the chart with respect to any one criterion, we can imagine the control lines to slide so that the median abscissa passes through the median mark of that class and read any significant deviations outside the lines in the new position.

Although analysis of variance has a similar function as the control chart, a control chart provides the same general results and with additional information.

Significance of Differences

If analysis of variance reveals significant variation between subclasses of a group, the calculation of the significance of differences between means of these subclasses analyses the case further.

The standard error of the difference between means of x and y is

$$
\begin{aligned}
SE^2_{m_x \sim m_y} &= \Sigma(m_x - m_y)^2/n \\
&= \Sigma m_x^2/n + \Sigma m_y^2/n - 2\Sigma m_x m_y/n \\
&= \sigma^2_{m_x} + \sigma^2_{m_y} - 2r_{m_x m_y}\sigma_{m_x}\sigma_{m_y} \\
\text{But} \quad r_{m_x m_y} &= r_{xy} \quad \text{(in population)} \quad (1) \\
&= r_{xy} \quad \text{(in sample - approximately)} \\
\therefore \quad SE^2_{m_x \sim m_y} &= \sigma^2_{m_x} + \sigma^2_{m_y} - 2r_{xy}\sigma_{m_x}\sigma_{m_y} \\
&= \sigma_x^2/n + \sigma_y^2/n - 2r\sigma_x\sigma_y/\sqrt{n_x n_y} \quad \text{(by (19))}.
\end{aligned}
$$

Then, assuming an approximately normal distribution for the difference, its significance can be tested by the ratio

$$\frac{m_x \sim m_y}{SE_{m_x \sim m_y}}$$, which is significant over 2.5.

To test the significance between the sample mean and the true mean of a population, when the standard deviation of the sample is s, we calculate $t = \frac{m - \mu\sqrt{n}}{s\sqrt{n/(n-1)}} = \frac{m - \mu}{s/\sqrt{n-1}}$

and refer to Student's tables for determination of significance.

---

(1) Proof in Peters and Van Voorhis, p. 161-2.

The standard error of difference between standard deviations is

$$SE_{s_x \sim s_y} = \sqrt{\sigma_{s_x}^2 + \sigma_{s_y}^2 - 2r_{s_x s_y}\sigma_{s_x}\sigma_{s_y}}$$

But $\quad r_{s_x s_y} = r_{xy}$ (1)

$$\therefore \quad SE_{s_x \sim s_y} = \sqrt{\sigma_{s_x}^2 + \sigma_{s_y}^2 - 2r_{xy}\sigma_{s_x}\sigma_{s_y}}$$

$$= \sqrt{\sigma_x^2/2n + \sigma_y^2/2n - 2r_{xy}\sigma_x\sigma_y/2\sqrt{n_x n_y}} \qquad \text{(by (20))}.$$

These methods are illustrated in chapter IV below.

---

(1) Proof in Peters and Van Voorhis, p. 180-2.

6. Other Methods

Secular Trend

This takes up the study of marks of a class or group of candidates on a specific examination over a period of successive years or terms to reveal any definite change (in the same direction) found in that general level of the set of marks - the level being conveniently represented by their mean.

If any appreciable trend is evident, it is most likely to be a straight line trend characterised by the equation

$$T = a + bx,$$

where T is the value of the mean,
x the time variable.

The constants a and b of this line are determined by fitting the line to the observed values over a specified range by the method of least squares [1].

The value of a trend line as applied to level of examinations is evident in cases where the difficulty of the examination and the standards of marking are kept constant with time and not made to depend on the performance of any particular group of students on any occasion; for an upward or downward trend of the mean indicates a corresponding change in the preparation of the candidates as a whole.

In Chapter IV a trend line is fitted to the means of matriculation mathematics examination marks in the last six years studied - and an upward trend is suspected.

Determining the Error of An Examiner

This may be done by comparing marks of more than one examiner given to the same scripts. If the standard error of each examiner's marks is determined [2], this record is kept for a revision of the same examiners' standards in future examination results.

---

(1) Appendix A, para 3.
(2) Rhodes' formula, p. 18 above.

Fiducial Limits

In a normal distribution whose mean μ is unknown, we can determine the range of possible values of μ for which the sample mean m is not significantly different.

Because the distribution of m is normal in a normal distribution, the relative deviation of m from its expected value,

$$m - \mu / \sigma_m = (m - \mu) \sqrt{n}/s ,$$

must be less numerically than 1.96 if m is not significantly different at the 5% level of probability.

The condition, which is $\frac{m - \mu\sqrt{n}}{s} < \pm 1.96$,

is equivalent to $m - 1.96\ s/\sqrt{n} < \mu < m + 1.96\ s/\sqrt{n}$.

Then $m \pm 1.96 s/\sqrt{n}$ are the 95% fiducial limits for μ. "They are the limits within which μ must lie in order that the observed sample mean should not be significant at the prescribed level of probability."[1]

In the marking of a paper by a group of examiners, if we consider the marks of different examiners as sample values, and the true mark as the true population mean, we can calculate the fiducial limits of this mean, and thence deduce the possible range of the true mark.

Illustration[2]

The marks given to candidate 31, Special Place Examination, English I, part I, were:

Examiner	Mark (x)	x - m	$(x - m)^2$
A	35	.2	.04
B	29	5.8	33.64
C	43	8.2	67.24
D	43	8.2	67.24
E	43	8.2	67.24
F	39	4.2	17.64
G	35	.2	.04
H	35	.2	.04
J	25	9.8	96.04
K	21	13.8	190.44
	348		539.60

$m = 34.8$, $s = 53.96$, $1.96\sqrt{s^2/n} = 4.6$ (approximately)
Fiducial limits are 30.2 and 39.4 (in the 30- class); range = 22 marks.

(1) Weatherburn, p. 122.
(2) data from the Marks of Examiners, p. 115.

Factor Analysis

If two measures, x and y, are composed each of two independent components -

$x = a + c$, $y = b + c$ (where a and b are also independent),

then the coefficient of correlation

$$r_{xy} = \Sigma xy/n\sigma_x\sigma_y = \sigma_c^2/\sigma_{c+a}\sigma_{c+b}$$
$$= \sigma_c^2/\sigma_{c+a}^2 \quad \text{(if } a = b\text{)},$$

so that r is that proportion of the total variance of the measures which is due to the common factor present in both of them, or as the proportion of overlapping.

If a table of intercorrelations, therefore, is available between a number of measures, it will be possible to detect common factors between them. Then, if the measures are different evaluations of the same individual examination by different examiners, the general common factor will be the true ability of that individual. Assuming that differences between examiners arise solely from weighting differences, we can calculate the weighting for each examiner, which is his correlation with the true mark, by use of the table of intercorrelations between examiners, and then obtain the true mark as the weighted average of all the examiners' marks [1].

Considering examiner i's mark x_{ij} for candidate j as compounded of weighted components of the general and specific factors, we can express it as

$$x_{ij} = \sum_{k=1}^{f} r_{ik}\, x_{kj} \quad , \qquad \text{I}$$

where f is the number of factors.

If M_1 represents the matrix of examiners' weightings for the different factors,[2] and M_2 the matrix of the candidates marks on the different factors, then (by I) the examiners' marks for the candidates are the elements of M_3, where $M_3 = M_1 \times M_2$. II

The correlation between the marks of examiner 1 and 2 is

$$r_{12} = 1/n \sum_{j=1}^{n} x_{1j}\, x_{2j} \quad ,^{(3)}$$

which indicates that a matrix of intercorrelations M_R can be constructed equivalent to the product of two other matrices, or

$$M_R = M_3 M_3' = M_1 M_2 M_2' M_1' = M_1 M_1' \quad \text{(by II)} \qquad \text{III}$$

---

(1) Burt - The Marks of Examiners.
(2) We need consider only one general factor here.
(3) Where the marks are expressed in standard measure.

An element of M_R is (by III),

$$r_{12} = r_{1g}\, r_{2g} + r_{1e_1}\, r_{2e_1} + r_{1e_2}\, r_{2e_2} \ldots\ldots\ldots \quad (1)$$
$$= r_{1g}\, r_{2g}$$

(assuming only one common factor, g)(2) IV

$$\therefore \quad \frac{r_{12}\, r_{13}}{r_{23}} = \frac{r_{1g}\, r_{2g}\, r_{1g}\, r_{3g}}{r_{2g}\, r_{3g}} = r_{1g}^2$$

$$\therefore \quad r_{1g} = \sqrt{r_{12}\, r_{13} / r_{23}}$$ if there are only 3 examiners.

In the case of more examiners, the mean of similar estimations of r is taken, and

$$r_{1g} = \sqrt{\frac{1}{\binom{m-1}{2}} \sum^{m-1} (r_{1i}\, r_{1j} / r_{ij})} \quad .$$

This is the weighting for examiner 1. Similar formulas hold for the other examiners. (3)

---

(1) where g stands for the general factor,
e for specific factors.

(2) To this condition ~~is~~ corresponds the rank 1 of the intercorrelation matrix.

(3) Burt goes on to analyse the common factors after eliminating the first. This requires the orthogonal transformation of the resulting matrix of intercorrelations.

Chapter IV

AN ANALYSIS OF MATRICULATION MATHEMATICS EXAMINATIONS OVER A PERIOD OF TEN YEARS

The Palestine Matriculation Examination is a public examination held annually by the Board of Higher Studies in Jerusalem in the three official languages (English, Arabic, Hebrew) for students of the secondary-school leaving stage; and secondary school children, as well as "unattached" [1] candidates from all over the country join in it. For this reason, the group which meets every year for the examination is very heterogeneous - receiving different instruction in different languages under widely differing systems of education. We may therefore consider it as a random sample from the whole population of those who are at that special educational level in the country [2].

At the beginning of the scholastic year the Board issues a syllabus of the examinations containing, in detail, all requirements in all subjects, so that the conditions affecting various standards of different institutions are equalised, and the candidates are all expected to be on an equal level of instruction - that required by the examination.

The Elementary Mathematics examination is held in two parts - paper I being on Arithmetic and Algebra, and paper II on plane Geometry and Trigonometry. The duration of the examination for both papers is the same, and in totalling the marks for the final average the scores on the two papers are given equal weight. Elementary Mathematics is one of the four compulsory subjects, so that all the candidates, whose number since 1941 has been ranging between 300 and 500 every year, have to take it - a fact which makes it favourable for the theory of random sampling. One further advantage in choosing this subject for study (an advantage discovered after the study began) is that the same two examiners have been marking the scripts during the whole period, with only the following exceptions:

In 1937 and 1938 examiner 2 was different,
In 1943 examiner 1 was different.

---

(1) A term used to refer to candidates who prepare for the examination through private study.
(2) It was later learnt that the bulk of the Jewish population prepare for a different public examination of a similar nature in Hebrew.

Notwithstanding any interesting facts it may reveal, this enquiry has been carried out mainly in application of some of the statistical methods surveyed in Chapter III, and is here reported as a further illustration of them. The analysis is divided into sections for easy reference.

Annual Fluctuations in Level and Variability

Matriculation is a type of examination which is not designed to fit the preparation of the candidates who take it in any particular year, but rather one which recognises a certain fixed level of achievement and expects candidates to work up to that level if they hope to pass; it is a standard selective examination[1]. Moreover, the candidate groups themselves, being random samples from the same normal population[2], tend to possess the same general type of achievement over all the years. Of the factors influencing the marking in the ten years under study, we can safely assume as constant the following:

form and difficulty of questions,
general level and scatter of candidates' achievement[3],
the marking of the examiners.

The accompanying time series shows the mean and standard deviation of the total marks in the subject for every particular year from 1937 to 1946, inclusive. A striking feature of this series is the wide range fluctuations in the mean in the years 1937-40. From 1940 to 1946 it seems as though there is a definite trend to gradual increase.

To test whether the difference between the means of 1940 and 1946 is statistically significant, we calculate the standard error of the difference, assuming normality in the two distributions. With subscripts representing the corresponding years, the standard error may be expressed as -

$$\sigma^2_{m_0 \sim m_6} = \frac{s_0^2}{N_0} + \frac{s_6^2}{N_6} + \frac{2r_{06}\, s_0\, s_6}{N_{06}} \quad [4].$$

Here, $m_0 \sim m_6 = 7.84$
$s_0^2 = 285.2$
$s_6^2 = 412.2$
$N_0 = 242$
$N_6 = 392$
$r = 0$, because the distributions are independent.

Then $\dfrac{m_0 - m_6}{\sigma_{m_0 \sim m_6}} = 7.84/\sqrt{1.18+1.05} = \underline{5.138}$, and the difference is significant.

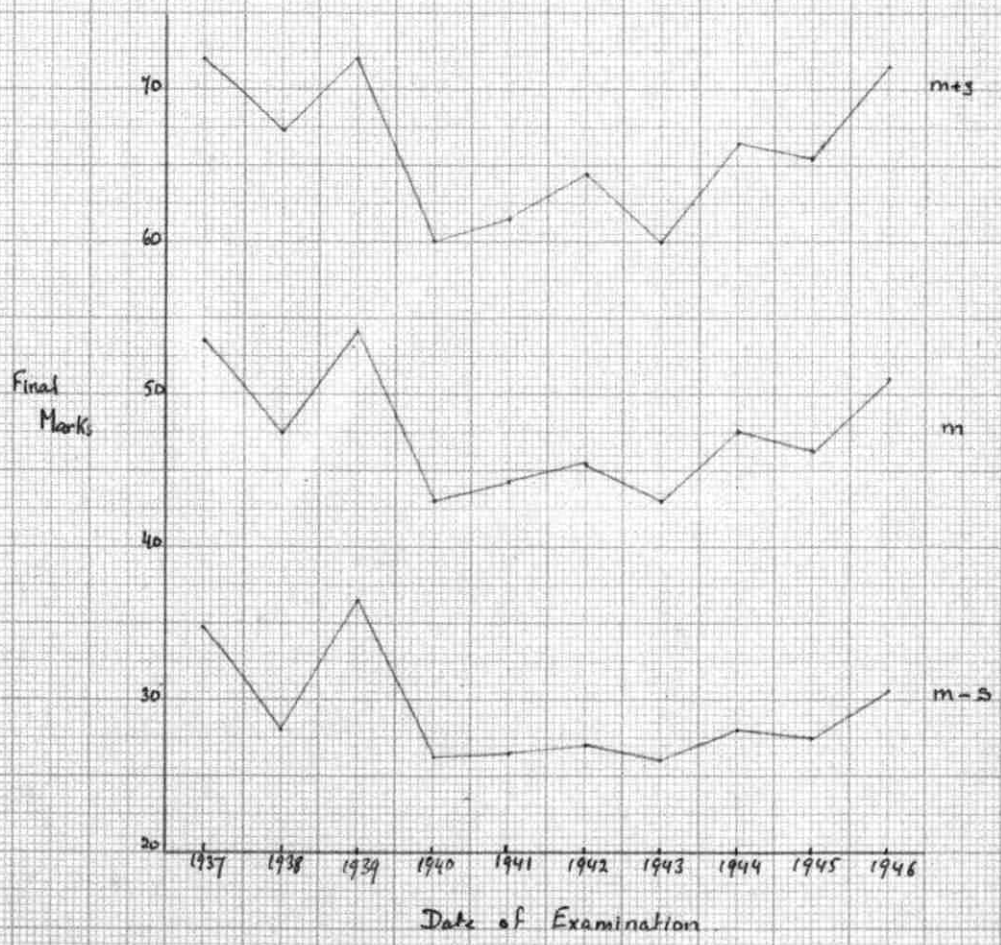
---

(1) cf. Ch. II for types of examinations.
(2) as deduced in para. 1 of this chapter.
(3) This hypothesis was later disproved by the method of standard error of differences. Actually, there are different schools that participate in the examination in different years.
(4) p. 74 above.

PALESTINE MATRICULATION –

MATHEMATICS.

Changes in Mean and Standard Deviation.



If we consider that section of the series bounded by these two years, we might suspect an upward trend of the mean. This cannot be decided without further data for later years. To measure the amount of this suspected trend, we can fit a trend line by the method of least squares, as follows -

The equation of the line will be

$$T = ax + b, \text{ where } \left.\begin{array}{l} a = \Sigma xy/\Sigma x^2 \\ b = \Sigma y/N, \end{array}\right\}^{(1)}$$

x maasures unit of time,
T is ordinate of trend line at x,
N is number of time units,
y is observed value of the measure at x.

For the data on hand,

X (year)	y (mean of marks)	x	x	xy
1940	43.2	-3	9	-129.66
1941	44.33	-2	4	- 88.66
1942	45.81	-1	1	- 45.81
1943	42.98	0	0	0
1944	47.77	1	1	47.77
1945	46.42	2	4	92.84
1946	51.06	3	9	153.18

$\therefore \quad b = 45.84,$
$\quad a = 29.66/28 = 1.06,$

and the trend equation is - $\underline{T = 1.06\ x + 45.84}$.

The chart does not show very high fluctuations in the standard deviation, but this can be judged by a further test of significance to test the difference between the highest and lowest values of the standard deviation - again for 1940 and 1946.

With the same assumptions, and using the same notation,

$$\sigma^2_{s_0 \sim s_6} = s_0^2/2N_0 + s_6^2/2N_6 \quad ^{(2)}.$$

Here, $s_0 \sim s_6 = 20.3 - 16.9 = 3.4$
$s_0^2 = 285.2$
$s_6^2 = 412.2$
$N_0 = 242$
$N_6 = 392.$

Then $s_0 \sim s_6/\sigma_{s_0 \sim s_6} = 3.4/1.056 = \underline{3.219}$.

---

(1) Appendix A, para 3.
(2) p. 75 above.

Referring to normal probability tables, we find that the probability of obtaining a deviation as great as this or greater is approximately 0.13 %, and the difference is therefore significant.

The null hypothesis, that the difference arises only out of random variation, is now disproved, and it appears that the two samples (the sets of marks in 1940 and 1946) cannot be regarded as coming from the same normal population.

Frequency Distributions

The conditions of setting the examination, the number of candidates taking it, the differences in their preparation - all are favourable to a normal distribution of marks for every year. However, some distributions have been found to give rather poor fits with the normal curve, and this can be inferred by an inspection of the frequency histograms for some of the years under study [1]. An actual fitting of the normal curve to the 1944 distribution was attempted, but the fit was not a very good one for, by the χ^2 test [2], it appears that the chances are slightly less than five to a hundred that it conforms with a normal distribution. The computations are summarised in a table as follows: [3]

x	z=(x-m)/s	½ erf z	½ Δerf z	½N Δerf z = T	Observed f. = O	$\frac{(O-T)^2}{T}$
0-	-2.422	-.5 (app)				
			.008	3	6	3
10-	-2.411	-.4920				
			.0713	27	24	0.33
20-	-1.408	-.4207				
			.1046	41	33	1.56
30-	-0.902	-.3159				
			.1623	64	55	1.27
40-	-0.395	-.1536				
			.1977	77	91	2.57
50-	0.111	.0441				
			.1852	72	64	0.89
60-	0.613	.2293				
			.1703	54	51	0.17
70-	1.125	.3696				
			.0778	30	37	1.63
80-	1.621	.4474				
			.0363	14	21	3.5
90-	2.138	.4837				
			.0163	6	6	0
						14.92

(1) Graphs for all frequency distributions pertaining to this study are found in Appendix C.

(2) p. 30 above. (3) Method and notation as in p. 59.

Then $\chi^2 = 14.92$. (degrees of freedom) = 10-3 = 7.

$m = 47.8$,
$s = 19.74$,
$N = 388$;

For 7 degrees of freedom, χ^2 at 5% level is 14.067,
at 1% level 18.475.

The frequency polygon (or histogram) for the 1939 distribution shows a negative skewness. We can obtain a rough estimate of this skewness by using the formula [1]

$$\text{skewness} = 3(m - \text{med})/s$$

Here, mean is at 54.2,
$s = 17.83$,
$\text{med at } 49.5 + \frac{40 \times 10}{56}$
$= 55.75$.

Therefore skewness
$= \frac{3 \times -1.55}{17.83}$
$= \underline{-.261}$.

Class	Cumulative Frequency
0-	6
10-	11
20-	35
30-	63
40-	108 — 148
50-	164
60-	229
70-	283
80-	293
90-	296

In the Instructions to Examiners it is pointed out that "with large numbers of candidates in a subject or in a group, a normal distribution of candidates in the subject or in the group should be approximately as follows:-

in group	A,	about	5%	of the	candidates,		
" "	B,	"	20%	"	"	"	
" "	C,	"	50%	"	"	"	
" "	D,	"	20%	"	"	"	
" "	E,	"	5%	"	"	"	.

This applies to candidates well prepared and to papers well balanced." [2]

In their system, the range of marks in percentage in the respective grades is as follows: [3]

---

(1) derived on p. 58.
(2) Part II, A, "Marking".
(3) Examiners are asked not to follow this rigid scheme blindly, but to vary the borderlines according to their own judgment and to the results of the class (by the methods of gaps).

86 - 100	in grade		A
66 - 85	"	"	B
46 - 65	"	"	C
31 - 45	"	"	D
0 - 30	"	"	E.

To examine the normality of this expected distribution, we can fit a normal curve by the usual method. In this case σ can be determined most easily by the quartiles, since

$\sigma = q/.6745$, where q is the semi-interquartile range.[1]

The distribution has quartiles at 45.5 and 65.5. The normal distribution should therefore have the mean at 55.5, and q equal to 10,

i.e. $\sigma = 10/.6745$,
$1/\sigma = .06745$.

Fitting a normal curve, using the above parameters,

x	$z = \frac{x - \mu}{\sigma}$	$\frac{1}{2}$ erf z	$\frac{1}{2}\Delta$erf z	Normal Scale $\frac{1}{2}\Delta$erf z x 100	Arbitrary Scale
0	- 3.743	-.5000			
			.0428	4	5
30	- 1.720	-.4572			
			.1968	20	20
49	- 0.708	-.2604			
			.4993	50	50
65	+ .641	.2389			
			.2378	24	20
85	1.989	.4767			
			.0233	2	5
100	3.002	.5000			

In practice, it was found that all the distributions in the results of the ten examinations deviated widely from this division[2] - although some were better approximations to the normal frequency curve fitted by means of the mean and standard deviation of sample - and the application of the χ^2 test proved the absence of any tendency to conform to any such an arbitrary distribution. The value of χ^2 for the 1946 comparison was 118.73, while 1945 gave 202.5 - as compared with the value for four degrees of freedom (in the five classes) of 13.28 on the 1% level.

---

(1) defined on pp. 26-7.
(2) In computing the frequencies in the respective grades, use was made of the frequency tables with class intervals of 10, and it was assumed that frequencies within a class are equally distributed on both sides of the mean. The errors involved in such an assumption are slight and do not account for the large discrepancies in the distributions.

Such large dis crepancies leave this arbitrary "normal" scale out of the question as a guide to examiners in the marking of these papers. Yet it was noticed that the frequency distributions for several years were very similar. In fact, after computing the average percentage of students in each grade over the ten years, the maximum deviations from the computed average were :

in the	A	grade	3	in	100	
" "	B	"	12	"	"	
" "	C	"	3	"	"	
" "	D	"	9	"	"	
" "	E	"	7	"	"	.

This suggests the possibility of fixing a "type distribution", for guidance in the marking of this particular examination, with a scale based on the calculated averages of the distributions in the ten years. The suggested type distribution will be --

	A	B	C	D	E
Percentage of frequencies (average of ten years)	3	19	34	26	18
Percentage of frequencies in Type Distribution	5	20	35	25	15
"Normal Distribution" (for comparison)	5	20	50	20	5 .

It obviously corrects for the tendency to have more E's, ~~a tendency well illustrated in the accompanying histogram.~~

There is little justification in using such a scale, since it is just as arbitrary as any other except that it has been based on the results of ten consecutive years, which should roughly represent the general tendency in such an examination. Theoretically, there is no reason for expecting such positive skewness, and the only argument in favour of using it in this particular case seems to be the fact that the majority of students are not on very good terms with mathematics - and it might be more repaying to seek normal achievement in the more popular subjects.

Papers I and II

The final examination mark is taken as the average mark of the candidate on the two papers. Although distributions in the totals marks have proved to deviate considerably from normal, frequency distributions on separate papers (1) may possibly still be normal. For, assuming normal

---

(1) Histograms in Appendix C.

distributions of marks given on papers I and II, their distribution functions would be -

$$dF_1 = e^{-\frac{x_1^2}{2}}\,dx_1\,,$$
$$dF_2 = e^{-\frac{x_2^2}{2}}\,dx_2\,,$$

and, for the distribution of their sum,

$$dF(z) = \int_{-\infty}^{\infty} e^{-\frac{(x-x_1)^2}{2}}\; e^{-\frac{x_1^2}{2}}\,,$$ (1) which is not a normal function.

Although in some cases (as in 1942, paper I) the distribution appears to be quite normal, most of the results have a detectable tendency to positive skewness, a property discovered in the totals distributions. A test of fitting the deduced "type distribution" of the totals to some separate paper distributions gave very good results in some cases, rather poor fits in others, for example

	I	II
1942	= 2.43, P .50	= 6.23, P .10
1945	=41.82, P .01	= 1.89, P .50
1946	=25.03, P .01	= 1.4 , P .50 .

The fitting of the same distributions to the expected normal scale gave almost as poor results as the fitting in the totals - being as high as 387.5(2) in 1945, paper I!

There is a general impression that geometry is more popular than the other branches of elementary mathematics. If that is so, students (generally speaking) tend to be better prepared in it. Is this the case ? and is general performance on paper II better ? (3) Granting that examiners' errors are equally effective in the two papers, a partial answer to this question can be obtained by comparing the means of papers I and II in several years.

The two sets of means are tabulated below, with the general means for comparison, and the significance of the differences between the means of the two papers tested for every year separately.

Year	Mean I	Mean II	G. Mean	M I ~ M II	σ^2_{MI-MII}	σ	MI ~ MII/σ
1942	45.15	47.84	45.81	2.69	1.86	1.36	2.5
1943	44.18	43.51	42.98	.67	1.51	1.23	2.5
1944	48.32	51.46	47.77	3.14	2.06	1.44	2.5
1945	43.47	50.66	46.42	7.19	2.22	1.49	4.82
1946	52.88	50.56	51.06	1.52	2.32	1.52	2.5

(1) Kendall, pp. 246-7.
(2) Comparing with 13.28 on the 1% level.
(3) On the other hand, it may be argued, geometry requires more specialised mathematical abilities, and is therefore more difficult.

It is obvious from the table that not only is the difference statistically significant, but that in the two cases 1943 and 1946 the mean mark for paper I is higher. The significant difference in the 1945 papers may be due to a specially difficult set of paper I questions.

Since both papers are set to examine the same function - mathematical knowledge and ability - it is also expected that their marks correlate rather highly, which can still be the case even if the difference in means is significant, since correlation measures relation between variates in standardised form (1). In terms of factor analysis, we can describe the correlation coefficient r between the two papers as an estimate of the common "mathematical factor" present in the two independent measurements.

The correlation coefficients calculated for the last five years were as follows :-

Year	r
1942	.781
1943	.755
1944	.7616
1945	.7433
1946	.7733 .

It is worth noting that the coefficients differ very little from year to year.

Examiners and Language Groups

The examinees may be divided into three groups (A,E, & H, say) according to the language in which they took the examination. The value of such a division lies mainly in the fact that the papers of the two groups A and H were corrected by different examiners, and an inspection of these two groups may throw light on their standards of marking.

A simple method is the testing of the significance of the difference, if any, between the means of the two groups. The means and variances of the marks of the two groups in 1941 were found to be (using the usual notation),

$$m_A = 45.17, \quad m_H = 35.03, \qquad \therefore m_A - m_H = 10.14;$$
$$s_A = 294.1, \quad s_H = 281.1;$$

Also, $n_A = 124, \quad n_H = 155.$

Then $\sigma_{M_A - M_H} = \sqrt{\frac{294.1}{124} + \frac{281.1}{155}} = 2.046, \quad \frac{m_A - m_H}{2.046} = \underline{4.95},$

a significant result.

---

(1) as deviations from their means and multiples of their standard deviations,

Tentatively, then, this cannot be a chance difference, and if further evidence is demanded for its existence similar tests can be easily carried out on the two groups in other years.

The difference may be due, as mentioned above, to the standards of marking of the two examiners, but another way of interpreting it would be the different level of achievement of students in the two groups of schools. A more comprehensive study of the situation is offered by Analysis of Variance - since it enables us to compare in this case the variance within a language group with the variance between means of language groups. Group E was included in the analysis [1].

The following table gives the frequency distribution of marks, subdivided into language groups.

		Number of Students			
Total Mark	x	A	E	H	Total Frequency
0-	-4	2	3	5	10
10-	-3	2	8	16	26
20-	-2	8	18	20	46
30-	-1	19	30	37	86
40-	0	27	37	31	95
50-	1	26	26	26	78
60-	2	20	11	16	47
70-	3	15	5	3	23
80-	4	4	5	0	9
90-	5	1	0	1	2
		124	143	155	422

As calculated, (in x units) $m_A = 0.67$, $m_E = -.13$, $m_H = -9.47$ Grand mean $M = -2.99$.

Correction term $= (\Sigma x_i)^2/N = 5 \times 2 + 4(9 - 10) + 3(23 - 26) + 2(47 - 46) + 1(78 - 86) /N.$

$= (10 - 4 - 9 + 2 - 8) /422$

$= 81/422.$

$\Sigma x_i^2 = 25 \times 2 + 16(9 + 10) + \ldots\ldots\ldots + (78 + 86) - 81/422$

$= 1331 - .2 = \underline{1330.8.}$

$\Sigma(m_j - M)^2 = 3.66 + 2.86 + 6.48 = \underline{63.56}$ [2]

$\Sigma_i \Sigma_j (x_{ij} - \bar{x}_{ij})^2 = 1330.8 - 63.56 = \underline{1266.36}$

Analysis of Variance

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares
Total	421	1330.8	3.16
Between means of groups	2	63.56	31.78
Within language groups	419	1266.36	3.022

(1) Papers in group E were corrected by both examiners.
(2) Subscripts: i stands for rows, j for columns.

$F = 31.78/3.022 = 10.52.$

In Snedecor's tables, for 2 and 400 degrees of freedom, respectively, the values of F are 3.02 at the 5% level;
4.66 at the 1% level.

The value of F obtained (10.52) exceeds these greatly, and we therefore conclude that the sampling into language groups is not homogeneous and that there is a significant difference between groups under this classification.

Control Charts

Analyses so far carried out dealt with scattered data from all over the range of period and is not reliable unless repeated on other samples chosen from the available data. In addition, the classification into language groups is inadequate and a further subdivision into centres of instruction (i.e. schools) can give better information.

A control chart[1] was thus drawn up covering four criteria of analysis:-

1) Language of examination (A, H, E)
2) Centres of instruction (A , E , H , i = 1, 2, ...)
3) Year of examination (1, 2, 3; i.e. 1941, 1942, 1943)
4) Papers I & II and total (× , ○ , ●).

The symbols inside the brackets are those used on the chart in various combinations, thus a compound symbol such as

● $2H_3$ represents mean of totals of centre H_3 in 1942,
○ $1E_{10}$ represents mean of paper II of centre E_{10} in 1941,
× 3A represents mean of paper I of language group A in 1943,
etc.....

The procedure followed in drawing up the chart was as follows:-

1) The marks in each of the years (1941, 1942, 1943) were divided into groups - every group representing one of the schools or centres of instruction. The mean mark for any particular centre in any particular year was then calculated by dividing the sum of marks for that centre by n, the number of candidates in it.

   This was done for paper I separately, paper II separately, and for the papers combined. Thus, for every centre four statistics were obtained -

   number of candidates,
   mean of marks on paper I,
   mean of marks on paper II, and
   mean of total marks.

   Centres with $n < 4$ were ignored in this chart.

---

(1) p. 73 above.

2) For the three language groups, in each of the three years, the same statistics were obtained [1].

3) Similarly, the general means on both papers and on totals were calculated for each of the three years.

4) The grand mean M for the whole period under discussion was found to be 44.3, and the standard deviation 15.66.

5) The standard deviation of the grand mean, s_M, is $s/\sqrt{N}$,
where s is the standard deviation of individual marks,
N the total number of candidates in the 3 years.

6) Assuming an approximately normal distribution, the majority of the means calculated in steps (1), (2), & (3) should fall within two "control lines" defined by the equations -

$$y = M \pm 2.5\, s_M \text{ [2]},$$

i.e. $$y = M \pm 2.5\, s/\sqrt{n} \qquad \text{(by step (5))}.$$

Substituting the statistics already obtained, the equations of of the control lines were found to be -

$$y = 44.3 \pm 15.66 \times 2.5/\sqrt{n}$$

or $$y = 44.3 \pm 39.15\, x \quad . \qquad (\text{where } x = 1/\sqrt{n} \text{ [3]})$$

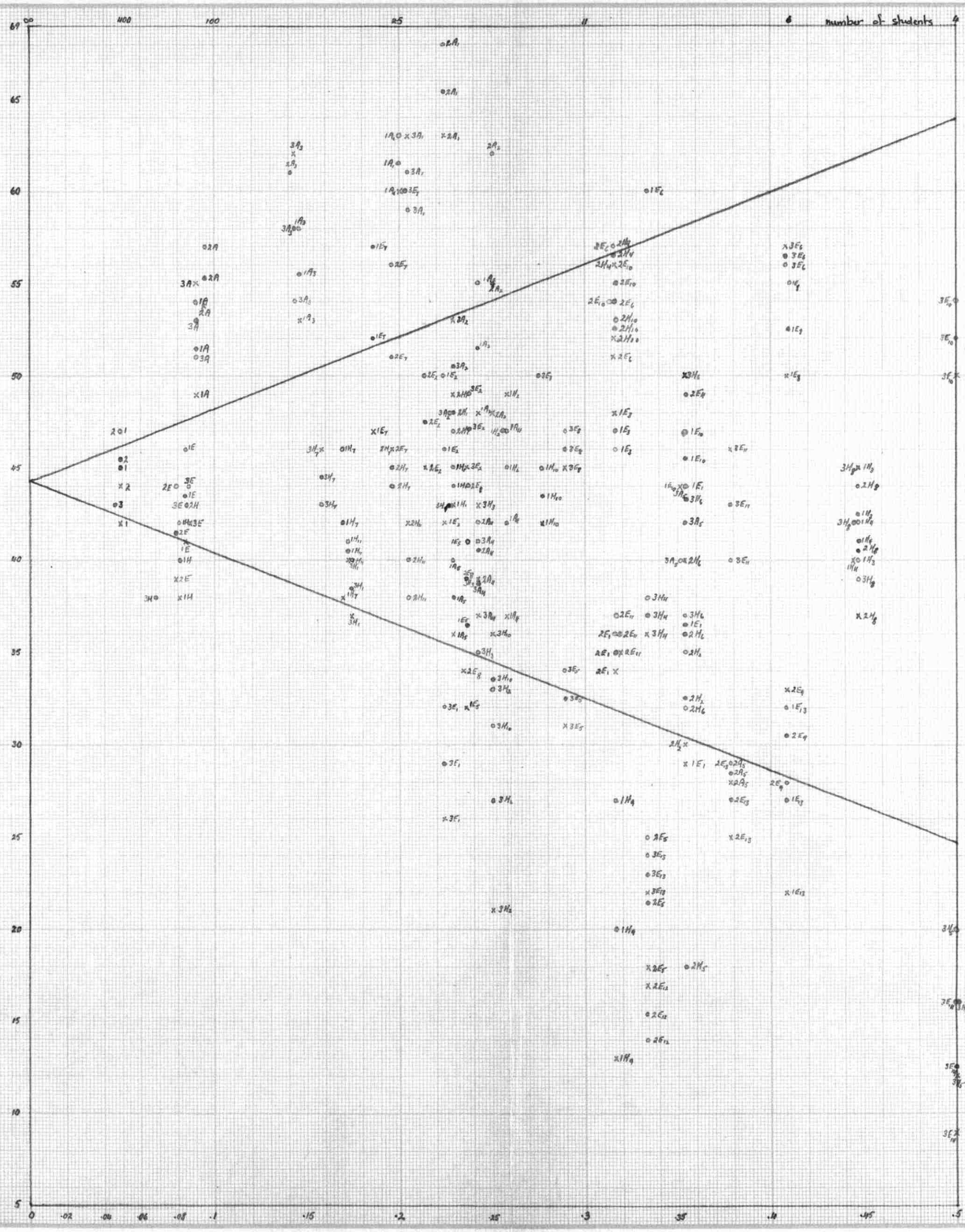
7) Taking y (or mean mark) as ordinate, x as abscissa with range .5 to 0 (i.e. n = 4 to ∞) [4], the two control lines were plotted. They meet at (0,M) where, n being infinite, the expected deviation of the mean is 0.

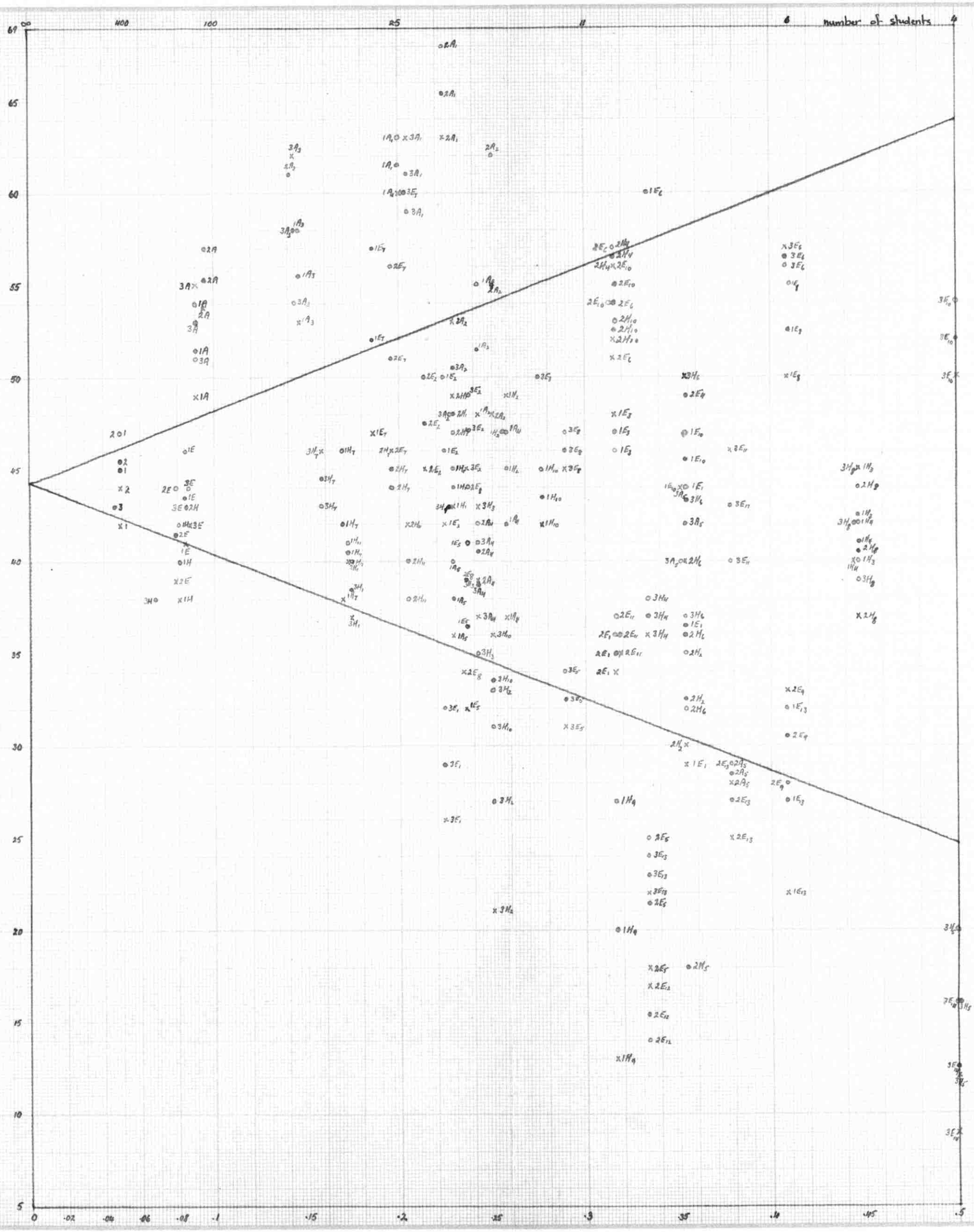
8) x (= $1/\sqrt{n}$) was calculated for each of the means in steps (1), (2), and (3) above.

9) With the x's as abscissae, and mean marks as ordinates, all the points were plotted [5].

---

(1) In calculating means for language groups containing centres ignored in step (1), account was taken of those centres and their contribution to the total.
(2) This was used, rather than the interdecile range suggested by Sandon, in order to facilitate computation, and because the annual means fluctuated only slightly.
(3) Here n is a variable because it assumes different values in different groups and centres.
(4) Note that centres with n 4 were not plotted.
(5) For guidance in interpreting results of graphing cf. pp. 14 and 74 above.


number of students
∞
400
100
25
11
6
4
69
65
60
55
50
45
40
35
30
25
20
15
10
5
0
.02
.04
.06
.08
.1
.15
.2
.25
.3
.35
.4
.45
.5


number of students
∞
400
100
25
11
6
4
69
65
60
55
50
45
40
35
30
25
20
15
10
5
0
.02
.04
.06
.08
.1
.15
.2
.25
.3
.35
.4
.45
.5

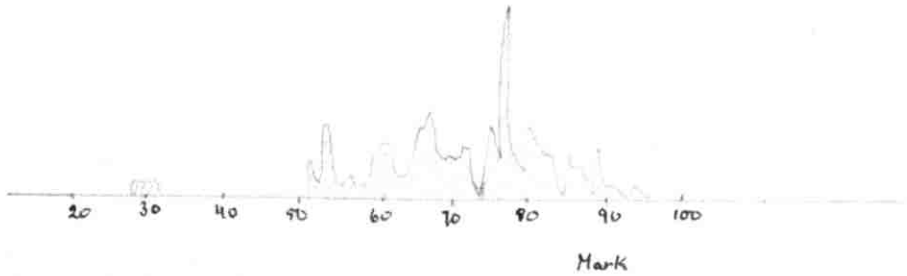
Chapter V

STANDARDISING MATHEMATICS EXAMINATIONS

Two reasons why mathematics will enjoy the privelege of a separate discussion are that, since all illustrations worked out came from mathematics examinations, the study naturally brought up points which required special attention; and more particularly, the ordinary traditional type of question found on a mathematics examination paper - the problem [1] - is set differently and requires different manipulation by the examinee than the other essay and objective type questions [2]. This applies equally well to questions in applied mathematics and to problem questions from the exact sciences.

Although mathematics is the most exact of the subjects on a school or college curriculum, examinations held in it require no less standardisation than the other examinations; and those who assert that by nature of its specifity questions and evaluation of answers are better standardised, appear to have as faulty an impression of the subject as those who claim that it is easier for any student to get high grades in mathematics examinations than elsewhere.

Among the investigations carried out by Starch was one on mathematics, and he concludes that mathematics papers are as open to unreliable marking as the more subjective subjects. 115 teachers corrected one paper of high school geometry, and the distribution of the results was as follows - [3]



(1) cf. classification of examinations, p. 15.
(2) An example of a geometry objective type examination is given in Wood's Measurement in Higher Education; but neither has this type of examination found wide acceptance yet, not is it exclusively objective - in the sense that the usual mathematical reasoning is involved in its s short answers. Besides, a student has also to be examined for ability to deal with complex situations.
(3) Described in Ruch, p. 49.

Hartog and Rhodes, in a similar investigation on 23 papers of University Honours Mathematics, found that there were more discrepancies than expected [1], the discrepancies being decidedly smaller when marking was done by pairing examiners, every two discussing papers together and marking according to a fixed scheme.

Needless to say, most of the methods already surveyed are just as applicable to mathematics examinations as to the other subjects. For instance, the expected form of the distribution of results may be a normal curve if the class is of a good size and ordinary ability [2]. Similarly, correlation of mathematics ability in school with intelligence may be a guide in the distribution of marks. Diagnostic curves and other graphs, as will appear below, can be of special value in this subject. Scores obtained on mathematics examinations can be converted into marks by some transformation [3], and distribution of results analysed by the usual methods.

Beside the use of ordinary methods, there are certain properties of the problem type of question that enter into the marking of these questions and are peculiar to it. For example, an examiner pays attention (in varying degrees) to method of attack as well as to the answer of the problem, to accuracy in calculation, logic in argument, form of answer, etc. In addition, if we wish to scale types of questions with regard to diagnostic ability, problems are an intermediate stage - since the student who has no idea about the solution gets no credit for guesswork or for "beating around the bush".

A student taking a mathematics examination is likely to come across problems which may be classified into three different kinds:

1) The type problem: which is not completely new to the average candidate. He probably finds himself familiar with the type, though not necessarily the details, and knows the general requirements for its solution through examples solved in text-books or common problems. But it takes a certain amount of work on his part to apply his previous knowledge to the new situation.

2) The challenging problem which, in contrast, presents very little familiar ground for comparison. The average student finds it somewhat far fetched. Rather than the usual routine methods, this problem usually requires an ingenious step to push the solution on - and the student is most likely to leave it till the end, if it is not already put there by the examiner.

---

(1) Table, p. 8.
(2) cf. results of enquiry in Matriculation examinations, p. 87.
(3) Corrected and original scores on Mid-year examination in Freshman, A.U.B., - on p. 46.

3) The third and most common type is the ordinary problem which demands direct steady attack (coupled with some patience at times) for its solution. Given the proper background of mathematical knowledge, the average candidate follows the usual methods and usually arrives (with luck) at the expected solution.

In view of the straightforwardness in its solution a problem of types (1) or (3) lends itself most to logical synthesis, and the person who sets the examination can classify the difficulty of a problem of this kind according to the number of independent steps in mathematical deduction required for its solution.

In solving a mathematical problem, the mind usually runs along lines such as the following:-

1. If it is merely a routine mechanical problem, one follows from the start the assigned plan of procedure.
2. If the problem is new, and not entirely routine work, the candidate reads it over until he understands its meaning and any technical terms involved.
3. He recognises to what branch of the subject it belongs, or knows the whereabouts of a possible solution.
4. He classifies the data, picking out what is relevant to the solution. Irrelevant data may mislead him.
5. He finds and establishes the relation between the data and the unknown.
6. If an auxiliary problem is offered, he can use it; if not, he searches for a plan.
7. He lays down the plan and follows it.

If we accept this scheme of solution as being typical of the average candidate confronted with a new mathematical problem of the ordinary type, then any such problem set on an examination may be analysed for the average student. Given the sufficient length of time, the student can tackle the problem. How much time is it right to give him?

It may be more convenient to regroup the above steps into separate independent sections, each section demanding a certain amount of time by itself for its performance. For the above thinking process, together with the necessary writing and calculation, the student will need -

(1) Time for recognising the problem (Recognition Time - steps 2 and 3 above): This involves understanding the problem and seeing its particular classification in the course. Technical terms add to the complication of the problem, and the time is therefore longer when more technical terms are included. With grade A students, or when the problem is familiar, this step is almost instantaneous.

(2) Time for expressing the problem in mathematical form (Translation Time - covering steps 4 and 5 above): This includes the drawing of figures in geometry, the setting of equations in algebra, etc. The student here has to differentiate between relevant and irrelevant data, and therefore the introduction of irrelevant data into the problem complicates this step of the solution.

(3) Time for planning a solution (Planning Time): This is the time it takes a student to work out a way of arriving at a solution. It is the most difficult part of the work and covers steps 6 and 7 in the scheme of thinking process.

(4) Time for dealing with routine operations, such as mechanical work, looking up in tables, simple algebraic or arithmetical manipulations, substitution in formulae, etc... This corresponds to step 1 above. If the problem is exclusively routine, this is all the time taken up in its solution [1].

In routine operations everyone has to follow the same method and go through all the necessary steps, and - except for the condensation in writing of two or three steps - there are no short cuts to the answer, and what the slower students have to do in writing, the quicker ones may work orally. It may therefore be possible to time such operations for any problem and evaluate how much time it takes an average student to go through the whole of the claculations.

Besides, these operations - being standard mental work - take as much time in one problem as in any other and instead of timing the amount of calculation in every problem set, we can resolve these calculations into elementary operations and time such operations separately - adding up different combinations of them in different problems. For example, we can time in arithmetic and algebra operations such as --

addition)
subtraction)
multiplication) specifying the number of digits
division) and number of terms.
raising to a power)
extracting a root)
removing brackets from an algebraic expression (per bracket)
plotting and graphing (number of points)
eliminating one unknown from two simple equations
factoring an expression (number of terms) .

---

(1) One or more of these "times" may not be required in some problems, for they may be already put in a simplified form suitable for students of a particular grade.

A brief Study of Trigonometry Problems

To test whether the above facts could be used to standardise the setting of mathematics problems, a brief study was carried out on problems of elementary plane trigonometry and although the results of the study can by no means be taken as conclusive evidence, yet they were encouraging and promise better success is larger scale investigations were undertaken on similar lines.

The scope of the present study was confined to the following types of problems :

1. Problems based on the solution of one right triangle.
2. Problems based on the solution of two right triangles.
3. Problems where further construction is required to reduce to one of the above two types.

The four divisions of solution time given in the general discussion above fitted the examples chosen in this study and the following list of routine operations (trigonometric, arithmetical and algebraic) was found to cover the necessary steps in section (4):-

1. Looking up a logarithm from 5-figure tables
   (a) with interpolation
   (b) without interpolation
2. Addition of two 5-figure logarithms
3. Subtraction of two five-figure logarithms
4. Looking up 5-figure trigonometric functions
   (a) with interpolation
   (b) without interpolation
5. Freehand construction of illustrative figures
6. Solution of a right triangle
7. Reducing algebraic expressions involving trigonometric relations
8. Factoring straightforward expressions
9. Solving a reduced simple equation of the first degree
10. Solving a quadratic equation by formula
11. Eliminating one unknown from 3 equations
12. Eliminating one unknown from 2 equations.

Eight persons,[1] including mathematics teachers and grade A mathematics students, were given a question paper [2] comprising operations and procedures in ordinary and type trigonometry problems in 4 sections:

---

(1) The small number is due to short time and limited scope.
(2) A copy of this is found in Appendix D.

Section I: Routine operations covering the above list. Candidates were asked to record (in seconds) how long it took them to go through each operation in writing at ordinary examination speed.

Section II: Problems to measure recognition time. Candidates were asked to record time taken in reading and understanding the problem.

Section III: Problems which require translation into mathematical form before solution. Some included various technical terms. Candidates were asked to read the problem and understand it first, then start timing for translation - thus eliminating recognition time.

Section IV: Problems which could not be solved at first sight, but which required some time for the planning of a solution. The candidates were asked to read the problem and record the time taken by them in attempting methods and working out the plans before arriving at the routine stage.

Since the planning of a solution may mix with the recognition or understanding of the problem, timing was done for the whole process of recognition, translation and planning, and then the same problem attempted after three days - a lapse short enough to keep the plan of the solution in memory but long enough too to eliminate any memory work in translation and reading. The difference between the two timings should give the time taken in planning in the first attempt.

The conformity in the results given independently by the different persons taking part (1) gave evidence to the correspondence of solution time taken by people of the same ability, and consequently to the possibility of standardising the optimum time for any rank of ability.

The following table gives the results of this brief study - the standard time taken as the average of the times given in the eight sets of results.

I. Routine Operations	Time
1. Looking up a 5-figure log, with interpolation	1′ 15″
without interpolation	30
2. Addition of two 5-figure logarithms	
3. Subtraction of two five figure logarithms	55
4. Looking up 5-figure trigonometric functions	
with interpolation	1 5
without interpolation	20
5. Freehand construction of figures (average of 3)	40
6. Solution of a right triangle (average of two)	3 30
7. Reducing algebraic expressions by trigonometric relations (time per step)	20
8. Factoring expressions (time per step)	1 15
9. Solving a reduced simple equation of first degree	35
10. Solving a quadratic by formula	1 35
11. Eliminating one unknown from 3 equations	2 10
12. Eliminating one unknown from 2 equations	1 10

(1) See Appendix D.

	Time
II. Recognition Time: (five of the candidates only)	20" - 45"
III. Translation Time: (five candidates only)	
Average time for translating one relation	15"
Average time for one mathematical term	30"
IV. Planning Time: (five candidates only)	
For one triangle problems	20"
For two triangle problems	40"
For problems requiring further construction	1'

For obtaining more reliable results, it is worth while carrying this out on a larger scale - with a larger number of testees, in any other branch, and on average students, so that the examination paper prepared by it is of the optimum difficulty.

Degree of Difficulty of a Problem

Still excluding from our discussion the challenging type of problem, we can adopt Burt's idea of "mental energy" as measurable by units of difficulty and rate of work. His exposition gives the physical analogy of work to the mental work involved in problem solution. He says:

> "We can then say that different tests [1], according to their difficulty, will require different amounts of the same '<u>mental energy</u>', irrespective of their specific quality or kind; and we can go on to declare that different individuals must possess a different amount of 'mental power', defining power in the usual way as the rate at which energy is expended, and measuring it in terms of the amount of work accomplished per unit of time." (2)

Burt's analogy runs as follows:- To describe any test performance we must state two things -

1. Number of problems solved in a given time ($d/t = v$),
2. Difficulty of the problem solved at a given average speed ($m.v/2$) - corresponding to $\frac{1}{2}$ the momentum.

And the amount of work done is mesured by the product $v.\frac{mv}{2} = \frac{1}{2} mv^2$.

---

(1) Here, "problems".
(2) The Factors of the Mind, p. 89.

One way, then, for estimating the degree of difficulty of an ordinary or type problem is by timing. For practical applications we may even express difficulty in terms of time units or equate a difficulty to a time unit. Thus a problem requiring 4.5 minutes for solution may be said to possess a difficulty of $4\frac{1}{2}$ units [1].

Illustration: Problem.

From a point on the ground at the foot of a column the angular elevation of the top of a tree is 68 degrees, and at the top of the column 30 feet from the ground it is 27 degrees. Find the height of the tree and its distance from the column.

Difficulty Analysis[2]	′	″
Recognition time		40
Construction of figure		40
Evaluating 2 trigonometric functions	2	10
Eliminating 1 unknown	1	10
Solving a simple equation		35
Translation: Two trigonometric relations		30
Planning 1[3] x 4	= 4	
	9	45

It is important to note the point where the greatest difficulty in solution appears. If it is at the beginning, the student may give up or waste most of his time without accomplishing any creditable portion of the problem. But if in a problem of equal difficulty units the obstacle is at or near the last stages of the solution, a student of the same ability is likely to accomplish more, and consequently receives more credit. It follows, then, that the first problem is actually more "difficult" - and to account for this different location of difficulty one way would be to weight planning time accordingly, giving it unity weight for difficulty at the end, 2 for difficulty in the middle, and four for initial difficulty.

In the above example, difficulty comes before translation, in the establishing of the two right triangles, a step which requires the construction of the dotted line in the figure. Difficulty units for planning are therefore 1 x 4 = 4.

---

(1) 1′ corresponding to 1 difficulty unit. This arbitrary relation may be easily altered by a change of scale.
(2) Estimates of difficulty are based on table of results, pp. 97-98.
(3) This is also based on the results of the study. Since there are types of planning procedures that hold good for certain groups of problems, it may be possible - having a sufficient collection of these "standard plans" - to determine the difficulty value of any one plan belonging to a particular type. The mathematics teacher can store a number of such plans (with their difficulty values determined) in every branch of his subject.

The total difficulty units taken up by the illustrative problem are $9\frac{3}{4}$, or 10 difficulty units. Thus, when setting this problem in an examination paper, the examiner can do one of two things:

Either (a) he sets the other questions with more or less equal difficulty,
or (b) he weights the scores on this problem with the ratio of 10 to the total difficulty units on the paper. Candidates must be warned of this, if there is a choice of problems.

Besides, the separate difficulty units assigned to the different steps in the solution may be used as a guide to the examiner in correcting the papers by his dividing the total credit on a question to parts corresponding to the separate steps, and then marking the papers with reference to this scheme.

On page 35 the results obtained by a class of 195 students on an examination were compared with the time taken by one examiner to solve the problems, and it was found that there was not the least correspondence in the results. But this should not be taken as evidence against the validity of evaluating difficulty by time, because the examiner should not be expected to work at the student average rate, and because those were results given by one examiner only, whereas the suggested method times as large a number of individuals as possible.

Marking Schemes for Mathematics Papers

In a standardised Reasoning Test, Monroe worked out a scheme for marking abilities in solving mathematics problems[1], and he based his marking on two scores which he called -

P scores, for solutions correct in principle,
C scores, for correct answers,

marking each problem for P first and, if correct in principle, for C also.

He also used a "rate score" to measure rate as well as accuracy of work - thus grading students for speed as well as method. The score he used is equivalent to the sum of P scores for all problems worked in 10 minutes (the students marking time on their papers at the end of every ten minutes). Obviously, he neglects accuracy in this.

Without necessarily adopting Monroe's marking methods, we can take account of the factors he introduced and list the various factors that should enter into the marking of a mathematics examination paper -

---

(1) Monroe, De Voss & Kelly, p. 62.

method
calculation
time
style and lay-out
ingenuity .....

If we use the analytic system and reduce every problem to its elementary operations and factors, the score is accordingly resolved into portions for method - in planning score
accuracy - in the calculation score
rate - in the total score, since the weighting for difficulty is based on time.

In marking mathematics papers, it will be useful to adopt a plan such as the following :

1) Set aside 5 - 10 % for lay-out.

2) Divide the total % into maximum marks for every question, proportional to their difficulty units.

3) Work out skeleton solutions for all problems, assigning a maximum for every separate unit in the solution. The weighting of the different parts of the solution depends on the objectives of the question.

Having fixed the plan, the teacher marks with reference to it. A detailed example from a non-mathematical examination is given in the next chapter.

Chapter VI

A PROPOSED SCHEME FOR STANDARDISING EXAMINATIONS IN AN INSTITUTION.

1. the Marking of an Examination

A large portion of the work in standardisation falls to the administration - but the examiner himself puts the initial steps and the degree of success of the administrative handling of the scores depends on the efforts of the individual examiner who has, in the first place, the important task of setting an examination with the proper (desirable) degree of difficulty, and then marking it as objectively as possible.

Having set a question paper of the proper degree of difficulty and with the best suitable diagnostic questions, the examiner starts preparing for the process of marking :

1) For each question he states the objectives in setting it - naming the factors that will affect his marking of the answer.

2) For each question he writes down a skeleton of the answer.

3) For each question he determines the degree of difficulty, and consequently its relative weight in the total.

4) Referring to 3), he divides the maximum total mark (say 100) into maximum marks for each question in direct proportion to their relative weights.

5) From 1), he draws up a list of primary, essential, independent factors (such as style, expression, method of attack, order, spelling,...) to cover those that enter into all the questions. These he calls F_1, F_2, ... $F_{\ell-1}$. The facts in 2) constitute the specific factor F_ℓ, which is different for each question.

6) He subdivides the maximum for each question Q_j into portions f_{ij} attributable to the different factors F (i = 1, 2, ... l). He is guided in this by 1) and 2).

7) He sets the elements obtained in 6) in a matrix of coefficients M_1 - the coefficients being the maximum marks assigned to every factor in every question[1]. For m questions, and l - 1 general factors and 1 specific factor,[2] the matrix will have m columns and m + l - 1 rows - the last constituting the diagonal matrix $[f_{lj}]$ (j = 1, 2, ... m).

$$\begin{matrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ f_{l-1,1} & f_{l-1,2} & \cdots & f_{l-1,m} \\ f_{l1} & 0 & \cdots & 0 \\ 0 & f_{l2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & f_{lm} \end{matrix}$$

8) In marking, the examiner prepares a second matrix M_2 whose elements x_{ij} show the proportion of the maximum which every candidate deserves on each of the l - 1 factors and on the lth factor in each question. For a class of n candidates, this matrix will have n rows and m + l - 1 columns. The entries in this matrix are made in fractions.

$$\begin{matrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1,l+m-1} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2,l+m-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & \cdots & x_{n,l+m-1} \end{matrix}$$

9) The marks of every candidate on each question are obtained by premultiplying M_1 by M_2[3]. The result is a matrix M_3 of n rows and m columns. Then the sum of the kth row in M_3 gives the mark of the candidate k on the whole examination.

$$\begin{matrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{matrix}$$

Using this method, the examiner has actually dissected every candidate's paper and divided it into a number of separate units - each unit being the candidate's credit in any one factor with prefixed weight; and in marking, his work is reduced to estimating what fraction of unity the candidate has succeeded in earning on each of the factors. The weights are the elements in the coefficient matrix M_1 and his fractions are the elements of M_2.

---

(1) <u>N.B.</u> Some of these elements may be zeros - if the factors are not involved in all the questions.

(2) Burt (The Marks of Examiners, p. 358) analyses the mark of an examiner given to a particular paper as depending on universal, particular (group) and specific factors. Here, the terms are applied to analyse factors in an examination paper, not in the conditions affecting the marking of that paper.

(3) i.e. the element in the pth row and qth column will be the inner product of the pth row in M_2 and qth column in M_1 -

$$x_{p1} f_{1q} + x_{p2} f_{2q} + \cdots\cdots\cdots\cdots + x_{p,l+m-1} f_{l+m-1,q}$$

This is candidate p's mark on question q.

An Illustration of the Application of Matrix Technique to the Marking of A High School Science Examination.

The illustration is taken from a question paper in general science set for a secondary class final examination[1]. The questions were :

1; Name the main classes into which foods are divided. Which of these do the following contain:
bread, eggs, beans, oranges ?
How would you test for them ?

2. Explain why vegetables and fruits are not very nutritious and why they are essential foods in spite of this fact.

3. What do you know about vitamins ? Mention four articles of diet, with the one or more vitamins it is believed they contain. What is the special value of each vitamin ?

4. What is the composition of ordinary soap ? Discuss the effects of soap on (a) hard water, and
(b) dirt of the skin during washing.

5. Make a clear diagram of a thermometer, explaining its construction carefully.

Knowing the required standards of the class, I drew up skeleton answers for all the questions and listed the factors that should affect the marking of each. The skeleton answers were essentially an enumeration of the facts, which I therefore took as the "specific factor" F_ℓ. A study of the questions and the expected answers resulted in the following analysis:

Question 1: Lay-out ;
Accuracy in scientific description (procedure of tests or experiments;
Number of facts = 5 + 4 + 5 = 14.

Question 2: Style - argument;
Accuracy in scientific description;
Number of facts = 4.

Question 3: Style;
Lay-out;
Number of facts = 12.

Question 4: Accuracy in description;
Use of scientific terms;
Number of facts = 2 + 1 + 1 = 4.

Question 5: Diagram;
Description;
Use of scientific terms;
Facts = 3 (essentials in the diagram).

---

(1) Jerusalem Girls College, College V, Domestic Science - June, 1941.

The questions are considered of equal difficulty, so that the maximum mark out of 100 allotted to each is 20. On the basis of this and the above analysis, M_1 was set out as follows

	Questions				
	Q_1	Q_2	Q_3	Q_4	Q_5
F_1 : Use of scientific Terms	0	0	0	4	2
F_2 : Accuracy in description	5	1	0	4	4
F_3 : Diagrams	0	0	0	0	8
F_4 : Style	0	3	0	0	0
F_5 : Lay-out, arrangement	1	0	2	0	0
F_6 : Facts	14	0	0	0	0
F_6 "	0	16	0	0	0
F_6 "	0	0	18	0	0
F_6 "	0	0	0	12	0
F_6 "	0	0	0	0	6
Maximum Mark for each question	20	20	20	20	20

This is the constant matrix, or matrix of coefficients.

For lack of sufficient material, M_2 was filled (with the exception of candidate C_2) for hypothetical candidates.

	F_1	F_2	F_3	F_4	F_5	F_6	F_6	F_6	F_6	F_6
C. 1	5/10	10/10	2/10	9/10	4/10	10/14	4/4	6/12	3/4	1/3
C. 2	0	4/10	9/10	10/10	8/10	8/14	3/4	12/12	4/4	1/3
C. 3	10/10	10/10	8/10	7/10	10/10	14/14	1/4	10/12	4/4	2/3
C. 4	8/10	8/10	4/10	5/10	2/10	6/14	2/4	11/12	2/4	2/3
C. 5	...	...	...	...	...	...	...	...	...	...

The entries for every candidate in each of the columns 1 - 5 are made at the end of the correction of his paper - after a second general look over it. They are put as fractions of 10 for convenience.

In the other columns entries are made as fractions of the total number of facts expected in the answers of each question - the numerator being the number of facts correctly stated by C. j, and the denominator the number of facts that should be correctly stated. The marks allotted here are therefore based essentially on counting, and need no subjective evaluation.

As can be easily verified by calculation, M_3 is the product of M_2 by M_1. For example, since the second row in M_2 and first column in M_1 are

	0	4/10	9/10	10/10	8/10	8/14	3/4	12/12	4/4	1/3
and	0	5	0	0	1	14	0	0	0	0 ,

row 2 and column 1 of M_3 is

$$(0 \times 0)+(5 \times 4/10)+ \ldots\ldots\ldots\ldots\ldots\ldots +(14 \times 8/14)+ 0$$
$$= \underline{10.8.}$$

	Q_1	Q_2	Q_3	Q_4	Q_5	Total
C. 1	15.4	19.7	9.8	15.0	8.6	68.5
C. 2	10.8	15.4	19.6	15.6	10.8	72.2
C. 3	20.0	7.1	17.0	20.0	14.4	78.5
C. 4	10.2	10.3	16.8	12.4	12.0	61.7
C. 5	..	..	..	..	..	..

In the end the total marks are rounded up to the nearest integer.

One student (of the required standard) was asked to "take the examination" and answer the questions, as in any ordinary examination. This was C. 2 in the above illustration, and the entries for her paper given in M_2 are the result of examiner D's corrections. Four other examiners were asked to correct the same paper - two of whom, E and F, were given matrix M_1 and asked to mark strictly in accordance with it; while the two others, G and H, were left at liberty to mark as they wished - only being informed that the questions carried equal weight.

The results were:

Examiner	Mark of C. 2
D	72
E	71
F	73
G	90
H	83 .

For comparison, D, E, and F's separate entries in M_2 for C. 2 are given below :

	F_1	F_2	F_3	F_4	F_5	F_6	F_6	F_6	F_6	F_6
D	0	4/10	9/10	10/10	8/10	8/14	3/4	12/12	4/4	1/3
E	2/10	5/10	7/10	9/10	7/10	8/14	3/4	12/12	4/4	1/3
F	5/10	5/10	8/10	8/10	8/10	8/14	3/4	12/12	4/4	1/3

In conclusion, the advantages of this technique may be summarised as follows:

1. By keeping one of the two multipliers in the product constant, it reduces variable errors considerably.

2. Analysis of a mark into component factors increases objectivity of scoring(1).

3. Since the evaluation is made in terms of fractions whose particular contribution to the total is not obvious at the moment of assessing the score, an examiner is less likely to add or deduct marks on any particular question subjectively (to make it a round figure, or by the effect of a good or bad impression in part of the paper, etc.)

4. When more than one examiner correct the same set of papers, this brings their standards of marking nearer together. The constant matrix for the examination may be set by their joint recommendations (2).

---

(1) A candidate's contributions in the "general" factors may vary from question to question. Even if this be the case, it will be more fair to the candidate if the examiner considers that factor in general performance on the whole paper.

(2) It is obvious that rows 1 to 5 of matrix M_1 of the above illustration are filled in subjectively - but with the stating of the objectives beforehand, this subjectivity is reduced to a minimum. On the other hand, part of M_2 (columns 6 to 10) is purely objective.

2. Standardising within a class.

In any one class, there may be a wide range of difficulty between subjects or a wide range of standards of different teachers. In classes where students select their courses, the result is a few subjects that are taboo to everyone but the adventurous student, and a few others that are taken solely for the sake of filling in schedules and raising the final average. In ordinary full-schedule classes, these become labelled "difficult" and "easy" subjects by the students, who come to know the characteristics of the distributions of results in a remarkably short length of time.

In standardising the marking within a class, we can either

(a) bring all marking to a comparable level, or
(b) correct a student's final average for the difficulty of the courses he takes.

And this may be effected as follows:

1. In the beginning of the scholastic year give every class general examinations in all subjects.

2. Mark all subjects in the same class by the plan proposed above - preferably constructing the constant matrices by the collaboration of all the teachers of the class.

3. Test the significance of differences between standard deviations of marks in the different subjects - and require any extreme examiner to adjust his dispersion of marks in future examinations.

4. Use either (a) analysis of variance, or
(b) a control chart ,
to reveal differences in the levels of marking in different subjects.
If no wide differences occur, the level of marking is already standardised. The examiners may then be asked to use their constant matrices as guides in constructing others for future examinations.

5. For examiners who deviate widely from the expected range, the coefficient (constant) matrices are re-examined with the object of detecting the source of error.
(a) If it can be corrected, the matrix is reconstructed and the examiner takes note of the correction for the marking of future examinations.
(b) If the error is the result of the candidates' work, <u>either</u> the difficulty of the subject is adjusted to fit the class level,
<u>or</u> note is taken of its relative difficulty.

6. In giving final marks (for term, semester, or year work) to a class where the significant difference in level has not been corrected, the difficulty ratio is added to the mark list of every candidate to recognise the relative quality of his work.

3. Standardising between classes.

For comparison between classes, uniformity of mean is necessary[1]. To keep this uniformity,

1. Calculate the grand mean for every class (i.e. the mean of all the subject means). In finding the grand mean, express every separate mean of a subject in standardised units as $m\sqrt{n}/s$ before adding, if the dispersions are dissimilar.

2. Test the significance of differences between means, or use a control chart to bring out any such differences. Then,
   (a) If any significant discrepancy is observed and it is due to the effect of a particular subject in the deviating class, whose level will be corrected, then no account is taken of it;
   (b) If a significant difference occurs which is not the result of any special difference in the separate subjects, the level of the odd class must be corrected.

3. A linear transformation of all scores in the odd class corrects for its difference of level.

4. Standardising with Time.

The analysis between and within classes made at the beginning of the year aims to standardise the marking in an institution for the whole year. Naturally, since the classes consist mainly of the same students in consecutive years, the general level will be expected to remain stable from year to year as well as throughout the first year.

1. Every class in any year may be compared in level (by the standard error of difference between means) with the class below it the previous year - since it is essentially the same group of students - either in general achievement or in particular subjects.

2. Throughout his work, every teacher can collect as many diagnostic questions as possible and a conference of teachers in every particular subject can decide on constant matrices for every one of the specific types of diagnostic questions, to be used whenever such questions are set.

3. In final examinations[2], where standards of difficulty are

---

(1) The dispersions are apt to be similar in different classes of the same size. Smaller dispersions in smaller classes may be expected and need not be corrected for.
(2) This applies to public examinations as well.

most stable with change in time, typical factors and weightings are set down for every kind of question to guide in the construction of the coefficients matrix in subsequent years.

Throughout this paper it has been tacitly assumed that any errors in the marks of examiners will arise out of factors over which, at the moment of preparing or marking, they have no control - which is too optimistic an assumption, for errors may often be due to conscious inaccuracy as well. Actually, in a few small examinations, unreliable marking is not a grave problem, but it is particularly when the examination takes on the decisive function of passing or failing, or affecting in any way a candidate's future career, that the question of marking becomes serious enough to require the application of these standardising methods.

In any case, the task of the statistician is to make such methods available - with the expectation that others will make good use of them.

Appendix A

MATHEMATICAL NOTES

1. Derivation of Spearman-Brown's Prophecy Formula.

Consider two arrays, one in an x and one in a y series[1], with a certain degree of correlation between them. Then,

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} \quad . \qquad (1)$$

Next, consider each array to be made up of the sum of N sets of similar[2] arrays, of n elements each, in each of the 2 series, that is,

$$x = x_1 + x_2 + \ldots\ldots\ldots + x_N,$$
$$y = y_1 + y_2 + \ldots\ldots\ldots + y_N.$$

Substituting in (1),

$$r_{xy} = \frac{\Sigma(x_1 + x_2 + \ldots x_N)(y_1 + y_2 + \ldots y_N)}{\sqrt{\Sigma(x_1 + x_2 \ldots x_N)^2 \Sigma(y_1 + y_2 \ldots y_N)^2}}$$

$$= \frac{\Sigma x_1 y_1 + \Sigma x_1 y_2 + \Sigma x_1 y_3 + \ldots\ldots \Sigma x_N y_N}{\sqrt{\Sigma(x_1 + x_2 \ldots x_N)^2 \Sigma(y_1 + y_2 \ldots y_N)^2}} \quad . \qquad (2)$$

Since the samples are similar, the standard deviations within either series are taken as equal. Dividing both numerator and denominator of (2) by $n\sigma_x\sigma_y$,

$$r_{xy} = \frac{r_{x_1y_1} + r_{x_1y_2} + \ldots\ldots r_{x_Ny_N}}{\sqrt{\underbrace{\sigma_{x_1}^2 + \sigma_{x_2}^2 + \ldots + \sigma_{x_N}^2}_{\sigma_x^2} + \frac{N^2 - N}{2} 2r_{xx}} \sqrt{\ldots}}$$

Assuming equal intercorrelations between sets within either series, the fraction reduces to

$$r_{xy} = \frac{N^2 \bar{r}_{xy}}{\sqrt{N + (N^2 - N) r_{xx}} \sqrt{N + (N^2 - N) r_{yy}}} \quad ,$$

where $\bar{r}$ is the average r .

---

(1) As, for example, marks given by examiners X and Y to a set of questions.
(2) To this condition corresponds, in application to examinations, the unity in material for questions and the uniformity in the choice of questions.

This is the generalised formula.

If x and y represent scores on the same examination, $r_{xx} = r_{yy}$, and the formula reduces to

$$r_{xy} = \frac{N^2 r_{xy}}{N + (N^2 - N) r_{xy}}$$

$$= \frac{Nr}{1 + (N - 1)r} \quad .$$

2. Rhodes' Derivation of an Examinaer's Deviation from the Ideal.

Let the marks given to a class of n candidates by m examiners be -

Examiners

		A	B	C	.....		m
	1	X_1	Y_1	Z_1	...	...	...
	2	X_2	Y_2	Z_2	...	...	...
Candidates	3	X_3	Y_3	Z_3	...	...	...
	.	...	...	...	...	...	...
	.	...	...	...	...	...	...
	n	X_n	Y_n	Z_n	...	...	...

For a candidate J, if the ideal mark is Q_j , then the marks he obtains under the different examiners are :-

$$\left.\begin{array}{l} X_j = Q_j + A_j \\ Y_j = Q_j + B_j \\ . \quad \ldots\ldots\ldots \end{array}\right\} \qquad (1)$$

We can consider A_j , the deviation of examiner A from the ideal mark, as consisting of two facotors $(\bar{A} + a_j)$ -

$\bar{A}$ being the average of his deviations for all the candidates, or $\sum_{j=1}^{n} A_j/n$ and denoting an estimate of his constant error, (2)

a_j his random variation for the marking of candidate J. His residual standard deviation of random variations over the whole group of scripts is

$$s_a = \sqrt{\sum_j a_j^2/n} \quad ; \qquad (3)$$

this indicates to what extent he introduces random variations in his marking from script to script.

If the above table of data be available, it will be possible to estimate the values of s_a, s_b, s_c, ... without knowledge of the ideal marks. The prodedure, as explained by Rhodes[1], is briefly as follows:

Let $\overline{X}$, $\overline{Y}$, $\overline{Z}$, ...$\overline{Q}$ be averages of the ma values of X_j, Y_j, Z_j, ...Q_j.

Take
$$\left.\begin{aligned} X_j &= \overline{X} + x_j \\ Q_j &= \overline{Q} + q_j \\ \cdot &= \cdots\cdots \end{aligned}\right\} \qquad (4)$$

Then, by definition,
$$\left.\begin{aligned} x_j &= q_j + a_j \\ y_j &= q_j + b_j \\ z_j &= q_j + c_j \\ \cdot &\quad \cdots\cdots \end{aligned}\right\} \qquad (5)$$

In (5), subtracting the second equation from the first, squaring, then summing over all the j's,

$$\Sigma(a_j^2) + \Sigma(b_j^2) - 2\Sigma(a_j b_j) = \Sigma(x_j - y_j)^2 .$$

Assuming the deviations of the markings of A and B to be independent, a_j and b_j are random, with equal probabilities of being positive or negative . The sum of their product should therefore vanish or be very small over the whole range of values.

Therefore, approximately, $\Sigma(a_j^2) + \Sigma(b_j^2) = \Sigma(x_j - y_j)^2$;
Similarly, $\Sigma(a_j^2) + \Sigma(c_j^2) = \Sigma(x_j - z_j)^2$;
... ...
... ...

If m is the number of examiners, there will be $m(m-1)/2$ such equations. Adding sets of these equations, we get -

$$\begin{array}{lllllll} (m-1)\Sigma(a_j^2) + & \Sigma(b_j)^2 + & \Sigma(c_j)^2 + \ldots & = & \Sigma(x_j - y_j)^2 & + \Sigma(x_j - z_j)^2 & \ldots \\ \Sigma(a_j)^2 + & (m-1)\Sigma(b_j)^2 + & \Sigma(c_j)^2 \ldots & = & \Sigma(y_j - x_j)^2 & + \Sigma(y_j - z_j)^2 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{array}$$

Putting
$$\left.\begin{array}{lll} \Sigma(a_j)^2 = \alpha & \Sigma(x_j^2) = \xi & \\ \Sigma(b_j)^2 = \beta & \Sigma(y_j^2) = \eta & x_j + y_j + \ldots\ldots = p_j \\ \Sigma(c_j^2) = \gamma & \Sigma(z_j^2) = \zeta & \Sigma(p_j^2) = 1, \\ \ldots & \ldots & \end{array}\right\} \qquad (6)$$

the equations may be written as

$$\begin{array}{l} (m-2)\alpha + \alpha + \beta + \gamma + \ldots\ldots\ldots = (m-2)\xi + \eta + \xi + \zeta \ldots\ldots - 2\Sigma(x_j p_j) \\ (m-2)\beta + \alpha + \beta + \gamma + \ldots\ldots\ldots = (m-2)\eta + \eta + \xi + \zeta \ldots\ldots - 2\Sigma(y_j p_j) \\ \ldots \quad \ldots\ldots \quad \ldots\ldots\ldots \quad \ldots \quad \ldots\ldots \quad \ldots\ldots \quad \ldots \end{array}$$

---

(1) The Marks of Examiners, pp. 189-193.

Adding, solving for $\alpha + \beta + \gamma + \delta \ldots$, and then solving for α, β, ... separately, we get

$$\alpha = \frac{m}{m-2}\xi - \frac{2}{m-2}\sum x_j p_j - \frac{1}{(m-1)(m-2)}(\xi + \eta + \zeta + \ldots) + \frac{1}{(m-1)(m-2)}1.$$

$$\beta = \frac{m}{m-2}\eta - \frac{2}{m-2}\sum y_j p_j - \frac{1}{(m-1)(m-2)}(\xi + \eta + \zeta + \ldots) + \frac{1}{(m-1((m-2)}1.$$

.. ..

The standard deviations are then readily calculated from these By the relations (6) and (3).

3. The equation of a Trend Line. (Fitted by Least Squares)

Let $T = a + bx$.

Then the condition is that the sum $\sum_i (y_i - T)^2$ is a minimum, or that its derivative is zero. The sum may be written as

$$(y - a - bx)^2.$$

Differentiating partially with respect to a and b, and equating the derivatives to zero,

we have $-2\sum(y - a - bx) = 0$
and $-2\sum x(y - a - bx) = 0$.

Then $\sum a + b\sum x = \sum y$,
and $a\sum x + b\sum x^2 = \sum xy$,
or $\sum x = 0$.

Therefore, $a = \sum y/n$,
$b = \sum xy/\sum x^2$.

4. An Estimate of Standard Deviation in a Grouped Normal Distribution.

In a normal distribution,

$$y = \frac{1}{\sigma\sqrt{2\pi}} \quad e^{-\frac{x^2}{2\sigma^2}}.$$

For a distribution with N_1 individuals,

$$Y = N_1 \, y.$$

Then

$$\begin{aligned}
\int_{-\infty}^{\infty} Y^2 \, dx &= \frac{N_1^2}{\sigma^2 2\pi} \int_{-\infty}^{\infty} e^{-x^2/\sigma^2} dx \\
&= \frac{N_1^2}{\sigma^2 2\pi} \times \frac{\sigma}{\sqrt{2}} \int_{-\infty}^{\infty} e^{-z^2/2} \, dz \\
&= \frac{N_1^2}{2\sqrt{2}\,\pi\sigma} \times \sqrt{2\pi} \\
&= \frac{N_1^2}{2\sigma\sqrt{\pi}} .
\end{aligned}$$

(If $z = \sqrt{2}\, x/\sigma$
($x = z\sigma/\sqrt{2}$
($dx = dz\,\sigma/\sqrt{2}.$

If N_2 is taken as sum of squared frequencies in intervals,

$$N_2 = \int_{-\infty}^{\infty} Y^2 \, dx = \frac{N_1^2}{2\sigma\sqrt{\pi}}$$

and

$$\sigma = \frac{N_1^2}{2\sqrt{\pi} N_2} .$$

5. Correction in r for Interval Grouping.

Let r_{xy} be correlation in terms of intervals,
$r_{x'y'}$ be correlation in terms of variates.

Now

$$r_{xy.x'} = \frac{r_{xy} - r_{xx'} r_{yx'}}{\sqrt{1-r_{yx'}^2}\sqrt{1-r_{xx'}^2}} .$$

But this is partial correlation with x' kept constant, so that x is also kept constant,

$$\therefore \quad r_{xy.x'} = 0$$

$$r_{xy} = r_{xx'} r_{yx'} \qquad (1)$$

Similarly,

$$r_{x'y.y'} = 0,$$

$$\therefore \quad r_{x'y} = r_{x'y'} r_{yy'} \qquad (2)$$

$$\therefore \quad r_{xy} = r_{xx'} r_{x'y'} r_{yy'} \qquad \text{(by (1))}$$

$$\therefore \quad r_{x'y'} = \frac{r_{xy}}{r_{xx'} r_{yy'}} . \qquad (3)$$

If $\bar{x}$ is the mean of x' in any column,

$$\bar{x}' = r_{xx'} \sigma_x/\sigma_x \; x \qquad \text{(assuming rectilinear regression of individual variates on class variate)}$$

But $\bar{x}' = x$, $\quad \therefore \; r_{xx'} = \sigma_x/\sigma_{x'}$

Substituting in (3),

$$r_{x'y'} = r_{xy} \frac{\sigma_{x'} \sigma_{y'}}{\sigma_x \sigma_y}$$

---

(1) Formula given in Kendall, p. 136.

6. A Correlation Ratio without Bias.

The correlation ratio $\eta^2 = 1 - \sigma_c^2/\sigma_y^2$.

Using estimates of population variance, s_c^2 and s_y^2 ,

$$\sigma_{cj}^2 = \frac{n_{cj}}{n_{cj}-1} s_{cj}^2 .$$

Summing and clearing from fractions,

$$\sum_1^k (n_{cj}-1)\sigma_{cj}^2 = \sum_1^k n_{cj} s_{cj}^2$$
$$(\Sigma n_c - k)\sigma_c^2 = \Sigma n_c s_c^2$$
$$\therefore \quad \sigma_c^2 = \Sigma n_c s_c^2/n-k$$

And unbiassed η^2 is
$$\varepsilon^2 = 1 - \frac{n_c s_c^2 (n-1)}{n s_y^2 (n-k)} \qquad (1)$$

But
$$\left.\begin{aligned} n_c s_c^2 &= \Sigma\Sigma (y_c - \bar{y}_c)^2 \\ n s_y^2 &= \Sigma (y - \bar{y})^2 \\ \eta^2 &= 1 - \frac{n_c s_c^2}{n s_y^2} \end{aligned}\right\} \qquad (2)$$

and
$$\varepsilon^2 = \frac{(n-1)\eta^2 - (k-1)}{n-k} .$$

7. Diagonal Method for Calculating r

Using subscripts for variables,

$$m_{x-y} = m_x - m_y$$
$$s_{x-y}^2 = \frac{\Sigma\Sigma (x-y)^2}{n} - (m_{x-y})^2$$
$$= \Sigma x^2/n + \Sigma y^2/n - 2\Sigma\Sigma xy/n - (m_x^2 + m_y^2 - 2m_x m_y)$$
$$= (\Sigma x^2/n - m_x^2) + (\frac{\Sigma y^2}{n} - m_y^2) - 2(\Sigma\Sigma xy/n - m_x m_y)$$
$$= s_x^2 + s_y^2 - 2r_{xy} s_x s_y .$$

Therefore
$$r_{xy} = \tfrac{1}{2} (s_x^2 + s_y^2 - s_{x-y}^2)/s_x s_y .$$

---

(1) Proof: Let x be deviation from sample mean,
x_t deviation from true mean.

$\Sigma x^2 = \Sigma x_t^2 - nm_i^2$, where m_i is mean in ith sample,
n is number of individuals in sample.

$\Sigma x_t^2 = \Sigma x^2 + nm'^2$

In k samples, $\sum_1^{kn} x_t^2/kn = \sum_1^k\sum_1^n x^2/kn + n\sum_1^k m_i^2/kn$

or $\sigma_x^2 = s_x^2 + \sigma_m^2$

$= s_x^2 + \sigma_x^2/n$

Therefore $\sigma_x^2 = s_x^2 n/(n-1)$.

APPENDIX B

Freshman Mathematics

American University of Beirut

1946 - 7.

Questions,

Frequency Distributions,

and

Diagnostic Curves for Questions

in Placement Test - October 1946,
First Quiz (Q) - November 1946,
Mid-Year Examination - February 1947.

Placement Test - Questions

1. A man buys 40 tons of a certain commodity at the price of LL 300 per ton. A year later he buys some more of it, obtaining now 30 % less of it than before but paying a price 35 % higher. By how much percent does the average price of the whole quantity that he bought exceed the original price (namely, LL. 300 per ton) ?

2. Simplify : $\dfrac{2x}{(5 - 4x - 2x^2)^3 \sqrt{(3 + 2x^2)}} + \dfrac{3(4 + 4x)\sqrt{(3 + 2x^2)}}{(5 - 4x - 2x^2)^4}$

3. The mid point of the sides BC, CA, and AB of a triangle ABC are D, E, and F respectively. The length of EF is 4 cms., the perimeter of the quadrilateral DEAF is 20 cms., and DE has two-thirds the length of BC. Find the length of the sides of the triangle ABC.

4. Prove: If the bisector of the angle BAC of a triangle meets BC in D, then BD:DC = BA:AC .
   Write down the converse of this proposition and state whether it is true or false.

Quiz I - Questions

Candidates are required to answer 3 of the following questions.

1. Express each of the following in its simplest form :-

   (a) $\dfrac{x^2 + 2}{x - 2} - \dfrac{(x - 2)(x + 2)}{x^3 - 2^3}$

   (b) $\dfrac{a^2 - a(a - b)/(1 + ab)}{1 - a(b - a)/(1 + ab)}$

   (c) $\sqrt[3]{4} \cdot \sqrt{3} + \sqrt[3]{3} \cdot \sqrt{2} \cdot \sqrt[6]{6}$

   (d) Simplify the following expression in such a way as to make its denominator real and rational :- $\dfrac{3 + 2\sqrt{-4}}{2 - \sqrt{-4}}$

2. A car goes 18 miles per hour faster than a truck and requires 3 hours less time to travel 240 miles. Find the rate of each.

3. (i) The mantissa of log 5 is 0.69897. What is the entire logarithm of
   (a) 500, (b) $\sqrt{50}$, (c) $\sqrt{0.05}$?
   (ii) Define the common antilogarith, and use your definition to simplify the expression

$$\frac{\text{antilog } (x + d) - \text{antilog } x}{\text{antilog } x}$$

   Hence find antilog 0.001, given that antilog 0.301 = 1.9999 and antilog 0.302 = 2.0045.

4. In North Carolina the average number N of inhabitants per square mile is given in the table below. Plot the graph. In what year was the population double that of 1840 ? What, approximately, was the population in 1915 per square mile ? (Use the graph for answering the questions above.

t:	1840	1860	1870	1880	1890	1900	1910
N:	15.5	20.4	22.0	28.7	33.2	38.9	45.3
t:	1920	1930					
N:	52.5	65.0					

Mid-Year Examination - Questions

Candidates are required to answer 5 of the following questions. They should note (i) that if they choose questions from Part I only they cannot score more than 90%, (ii) that the questions in Part II are somewhat harder than those in Part I.

Part I: Maximum 18 % for each question

1. (i) Reduce the following to its lowest terms

$$\frac{x^3 - 2x^2y + 4xy^2 - 8y^3}{16y^4 - x^4}$$

(ii) Simplify $\dfrac{\sqrt{x^{-1}y} + \sqrt{xy^{-1}}}{\sqrt{x^{-2}y^2} - \sqrt[3]{(xy^{-1})^4}}$

(iii) Show that both the sum and the product of the complex numbers $a + bi$ and $a - bi$ are real. (It is supposed that a and b are real)

2. (i) Find the common logarith, of 5 if that of 2 is 0.30103.
(ii) Using logarithms calculate the square root of

$$\frac{101.32 \times 0.0075625}{201.56}$$

as accurately as possible, and then round off your answer to 4 significant figures.

3. The height of a pyramid whose base is an equilateral triangle varies directly as the volume and inversely as the square of the side of the base. If the volume were increased by 21.5% and the side of the base were decreased by 10%, find the percentage increase in the height.

4. (i) If $y = 27x - x^3$ find the average rate of change of y with respect to x as x increases from -2 to 1.
(ii) Find approximately the distance travelled by a body whose motion is governed by the equation $S = t^2 + 10$, where t is in minutes and S in feet, during the first 12 minutes.

5. The vertices of a triangle are at A(0,0), B(a,0), and C(b,c).
(i) Find the equations of the sides.
(ii) Show that the perpendiculars from the vertices onto the sides opposite them meet in a point, and give the coordinates of this point.

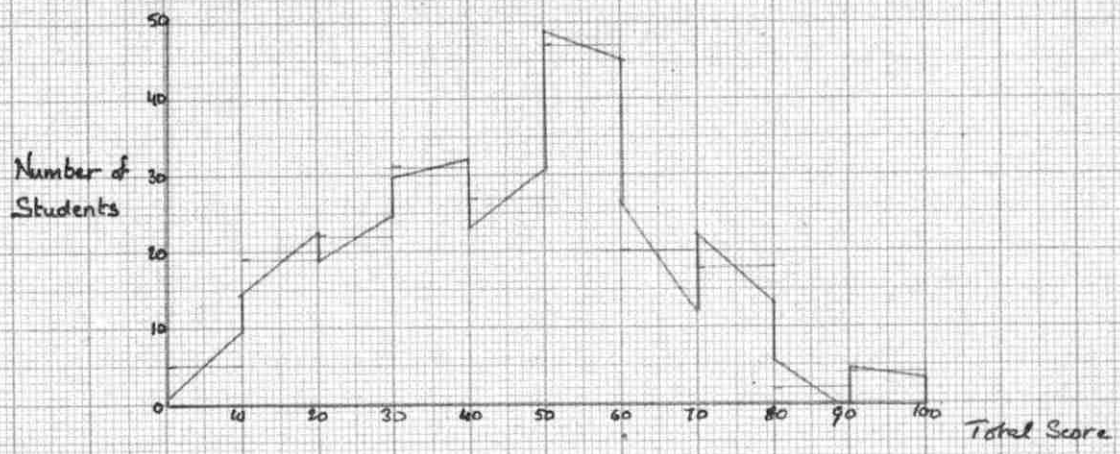
Part II: Maximum 23% for each question

6. Two trains start from A and B at the same time and, travelling towards each other, meet at C. Then the train that has come from A continues its journey to B, covering the stretch CB in 8 hours; and the train that has come from B continues its journey to A, covering the ~~whole journey~~ stretch CA in 2 hours. How many hours did each train take to make the whole journey between A and B ? (Assume that the speed of each train is constant throughout its journey).

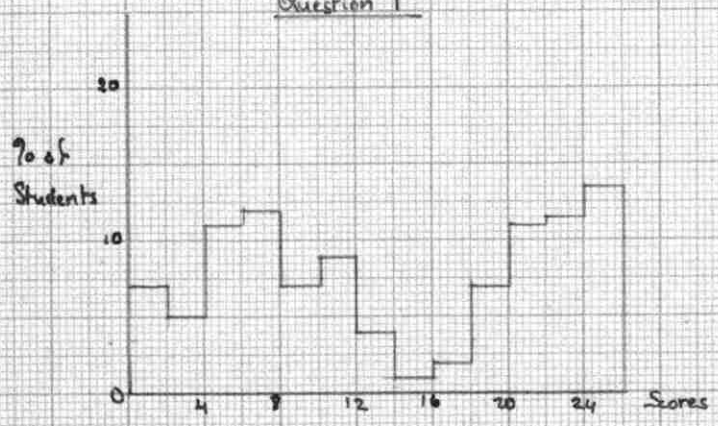
7. Two opposite vertices of a parallelogram ABCD are at A(-4,6) and C (8, -10). If the slope of AB is m find the equations of two sides.
   Find the value of m for which the parallelogram is a square and the values of the coordinates of B and D in this case.

Placement Test

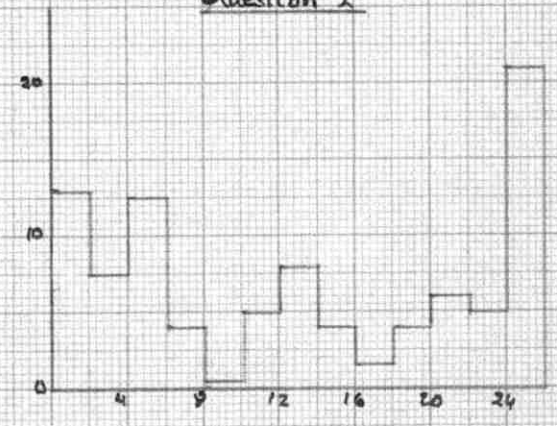
Frequency Distributions



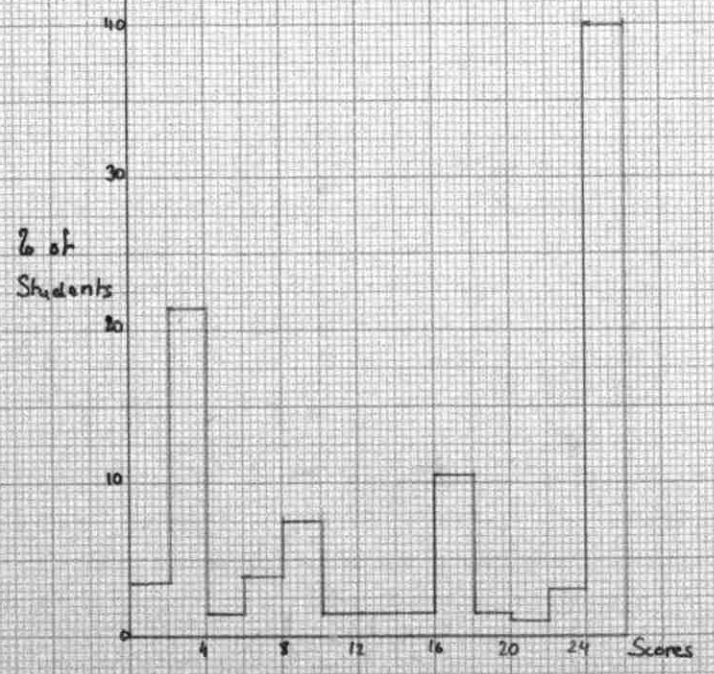
Question 1



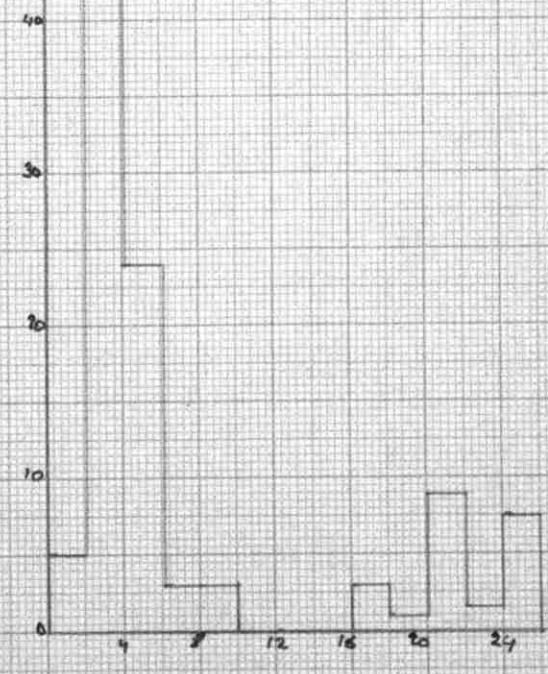
Question 2



Question 3

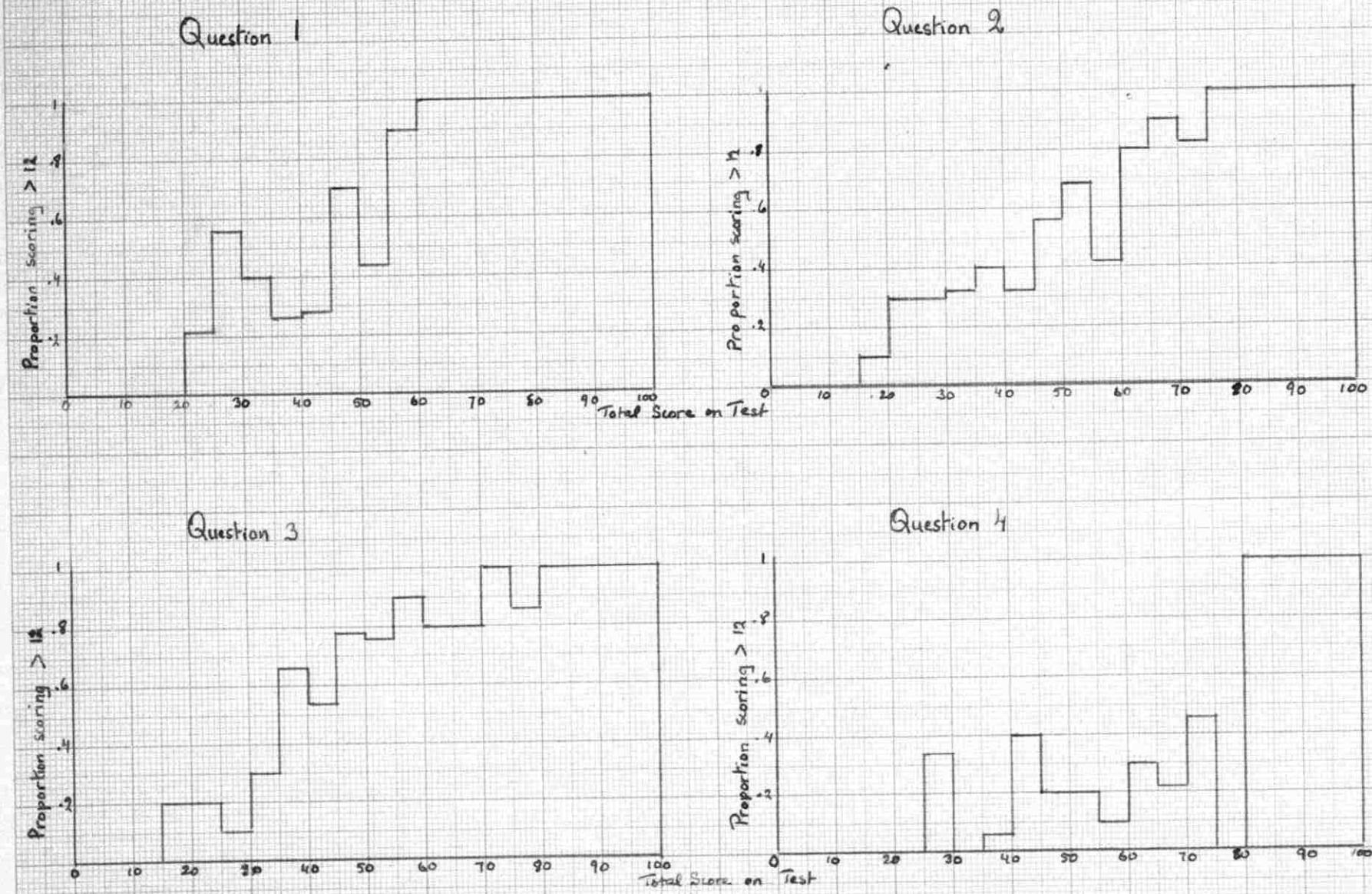


Question 4



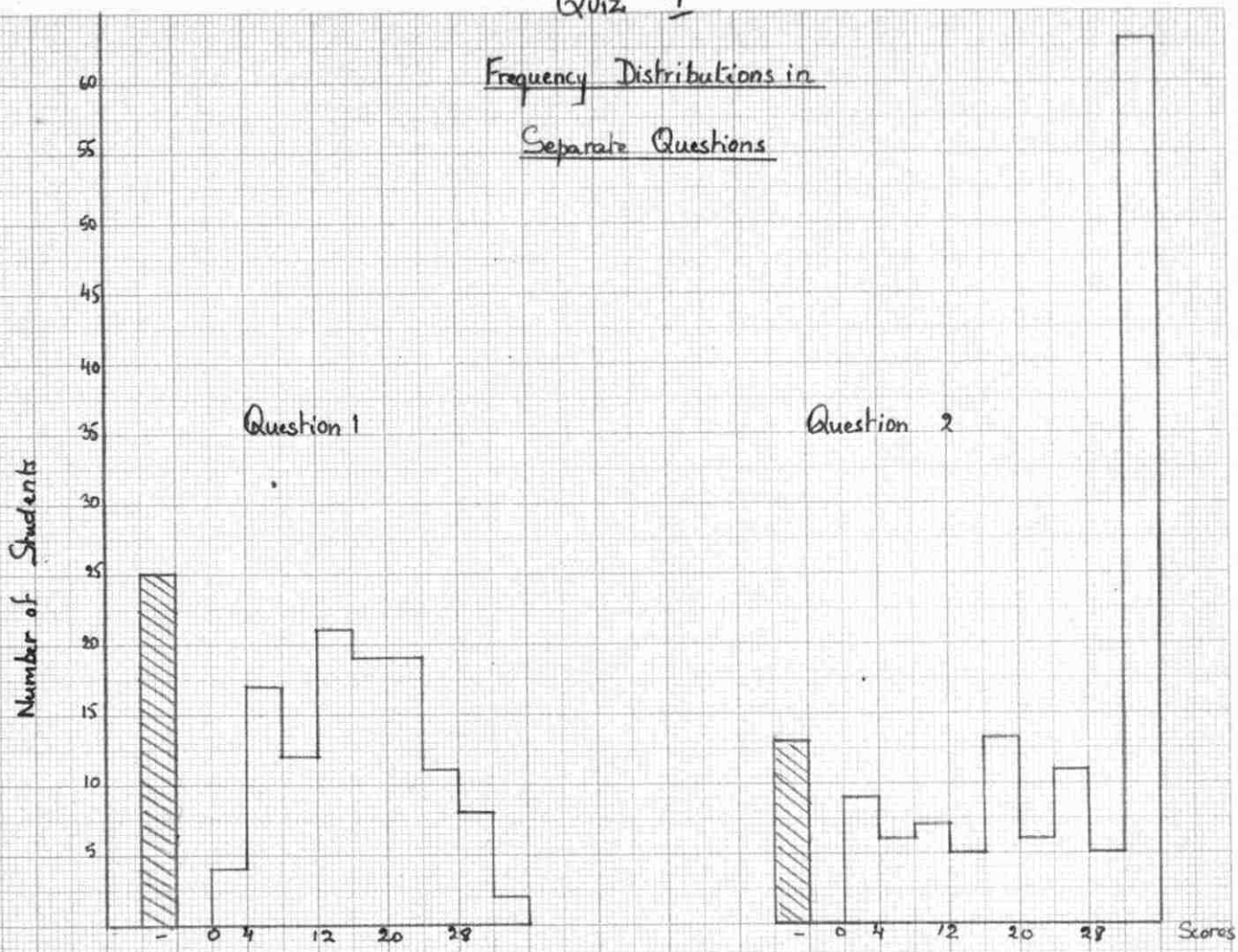
PLACEMENT TEST

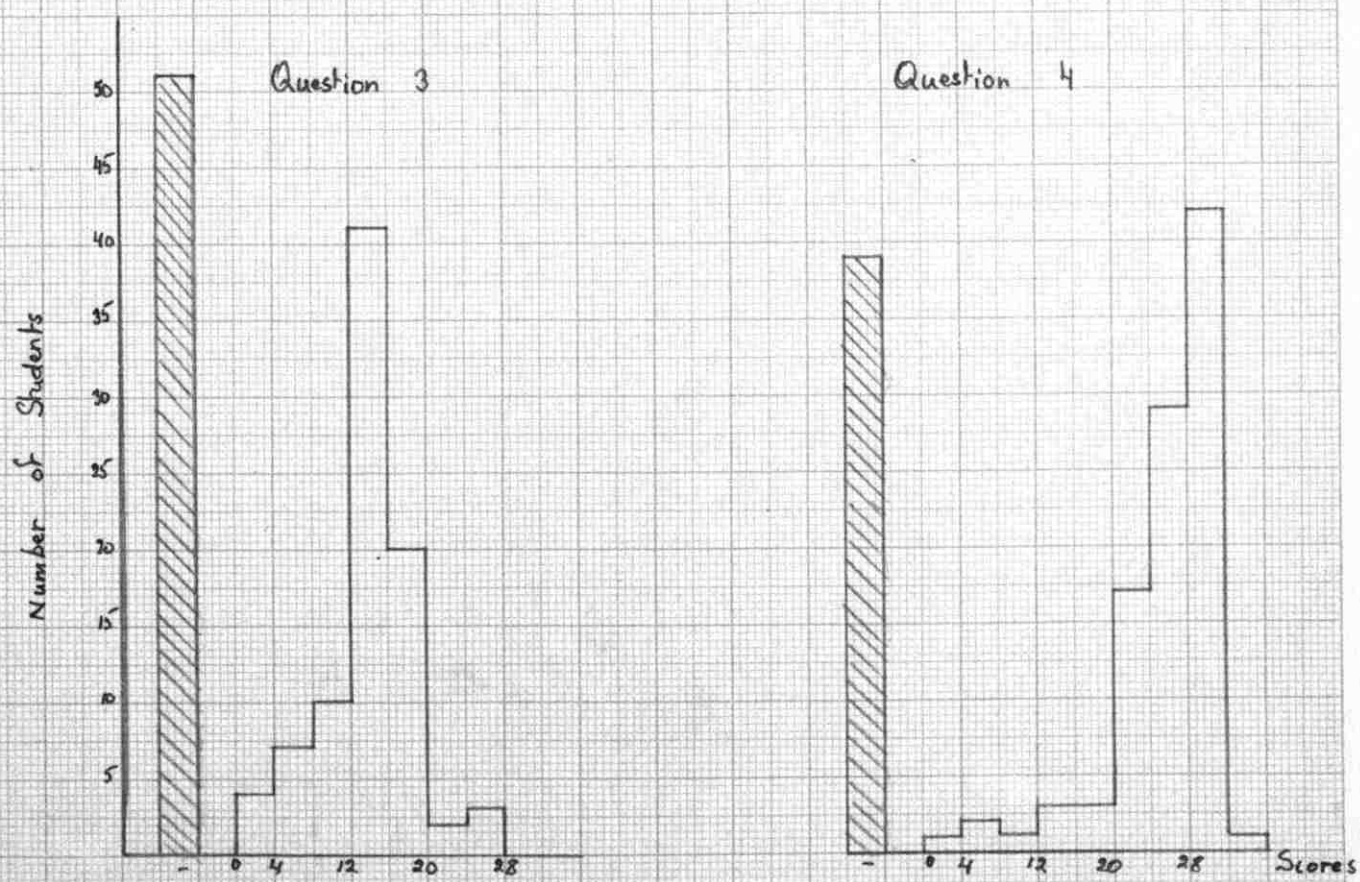
Diagnostic Curves



Quiz I

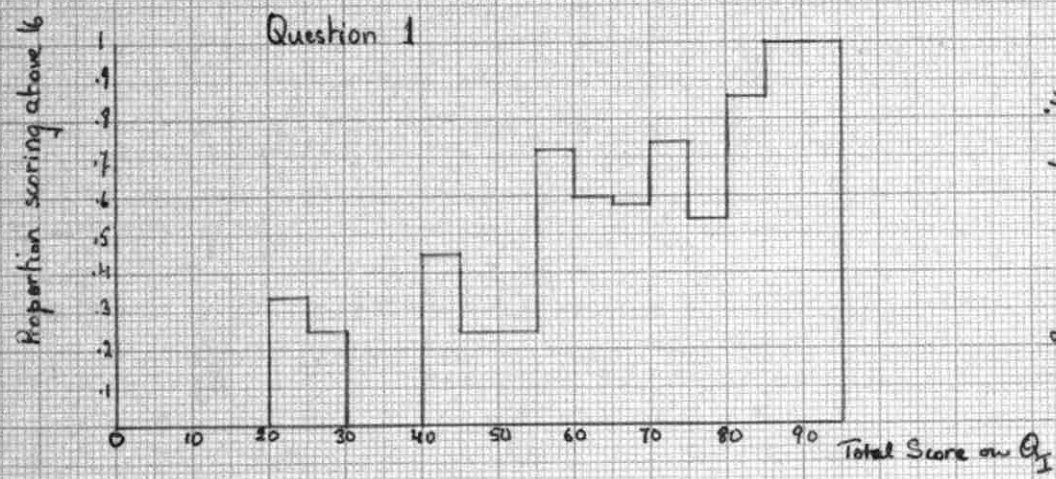
Frequency Distributions in Separate Questions

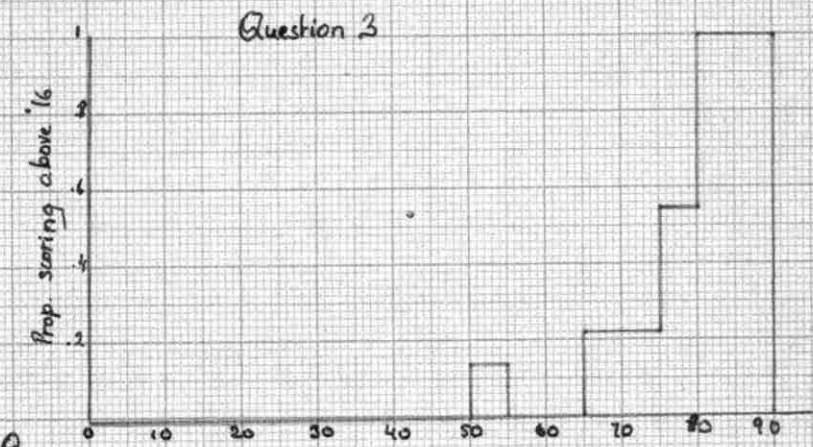


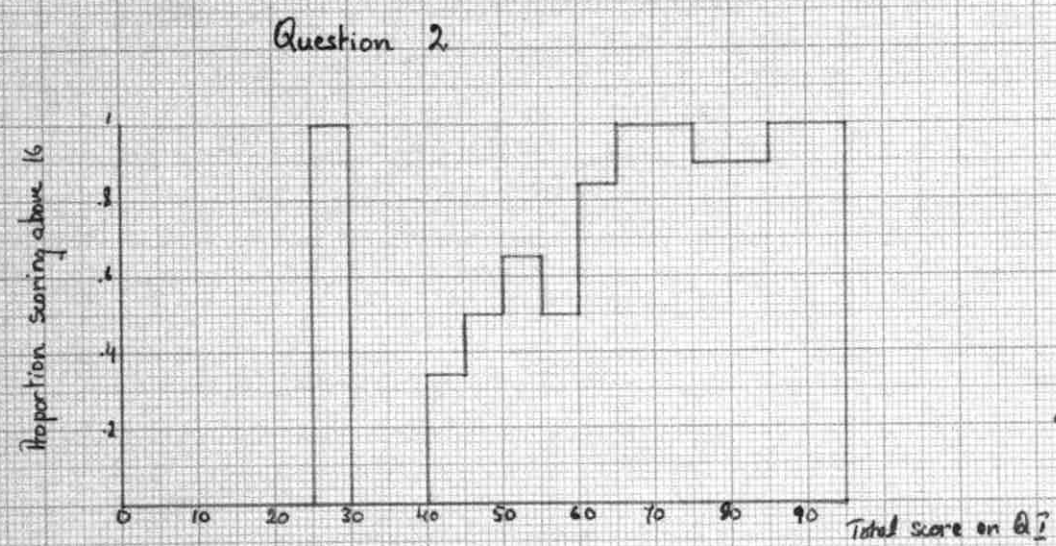


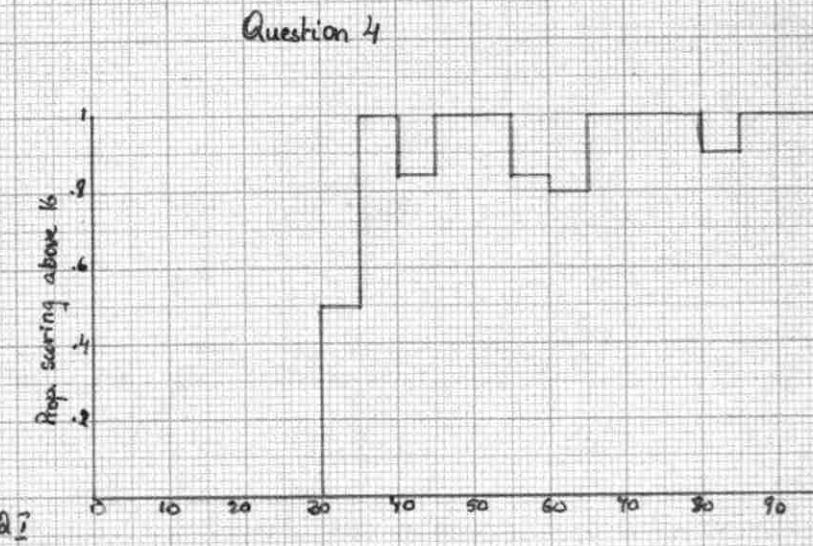
Quiz 1

Diagnostic Curves

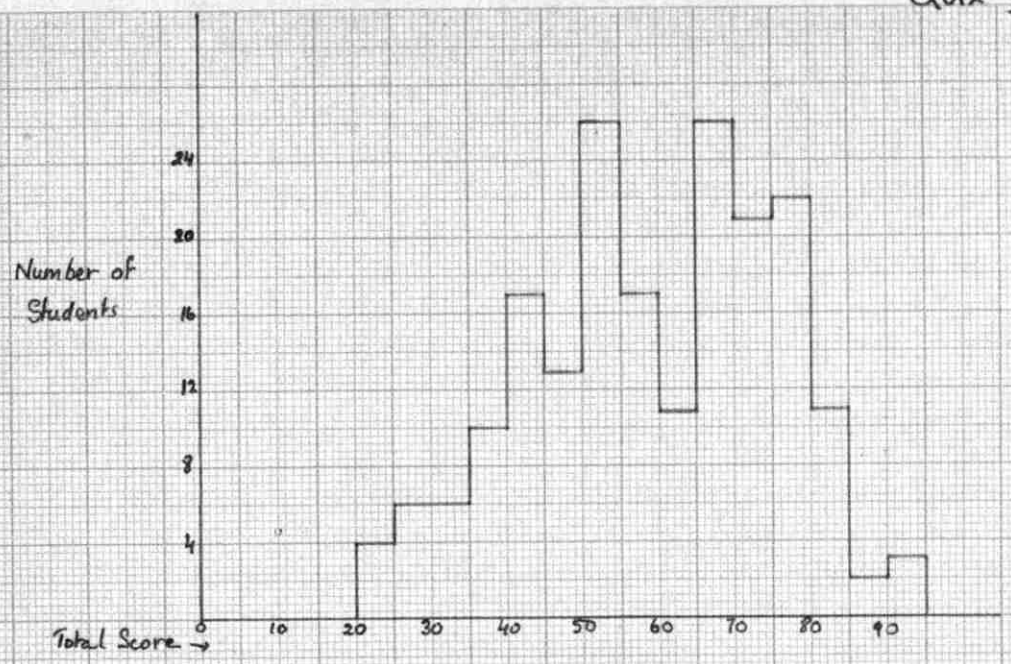




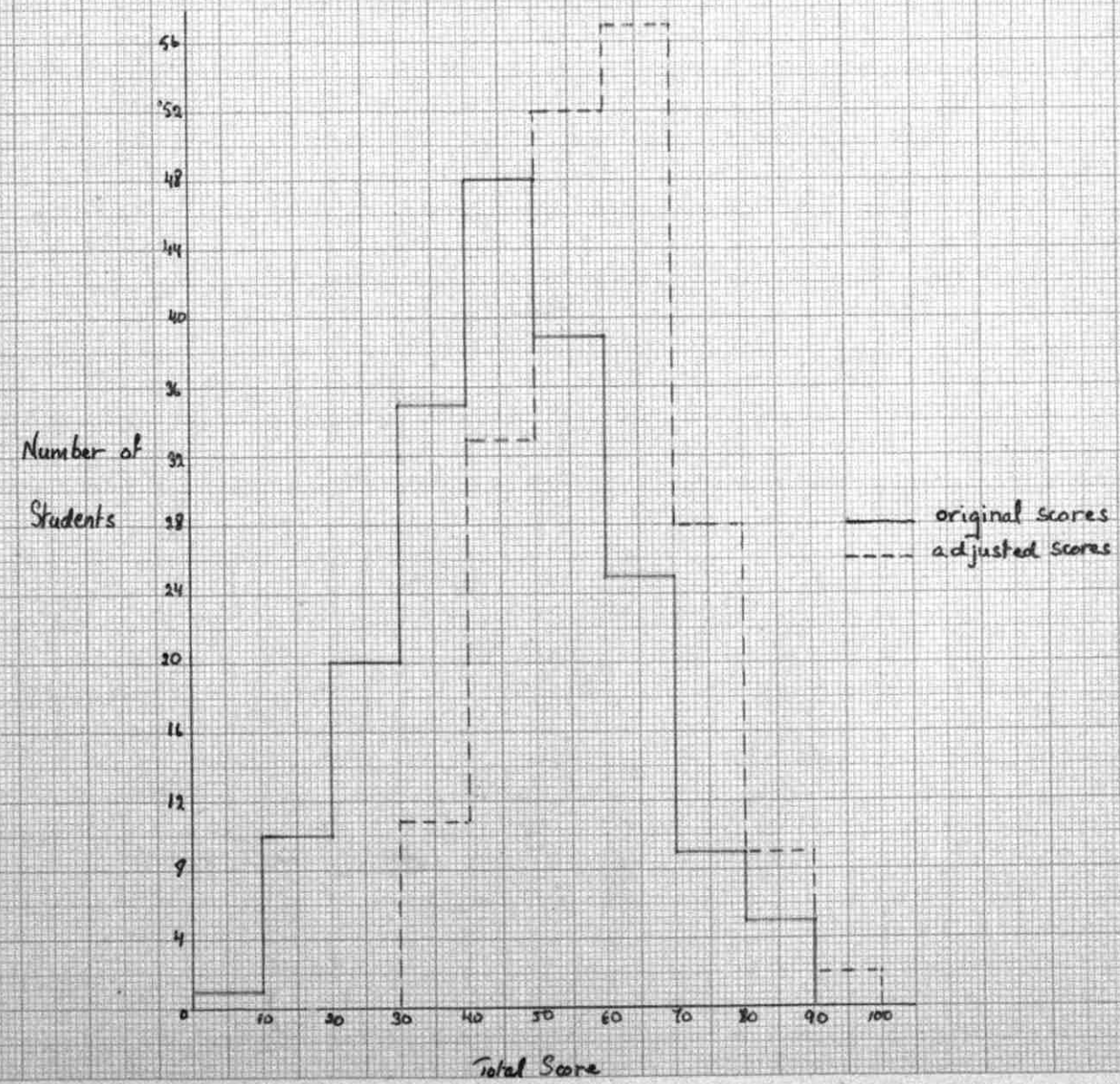




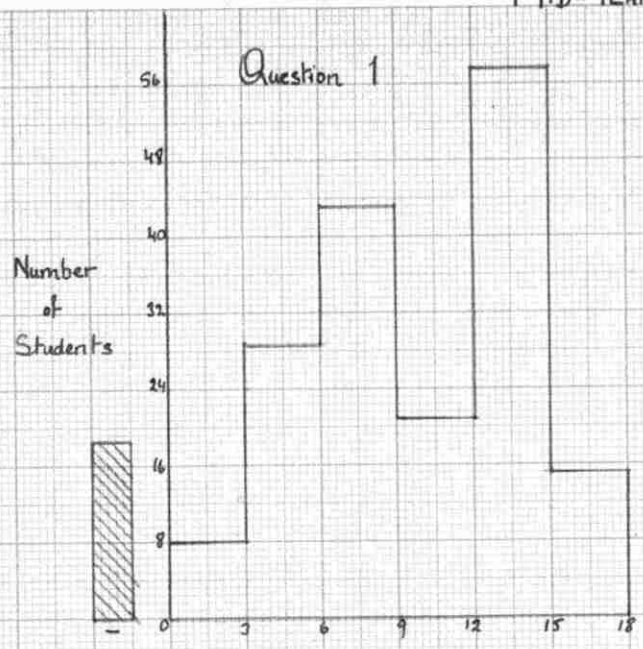
Quiz 1

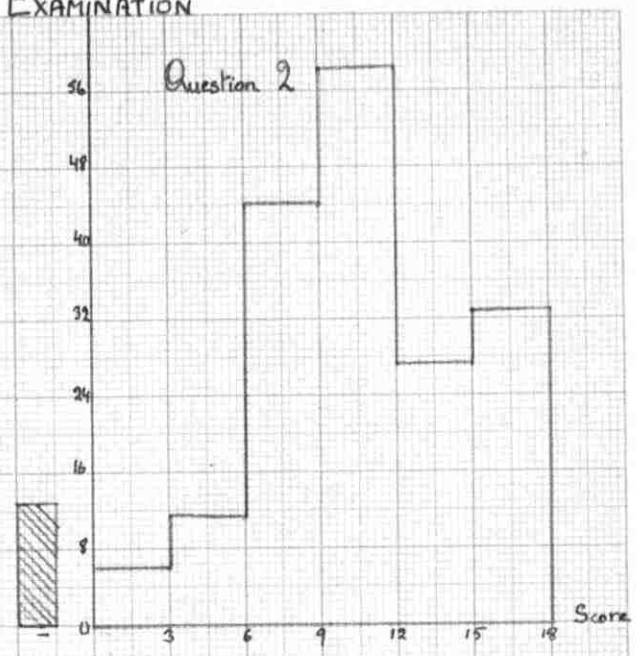


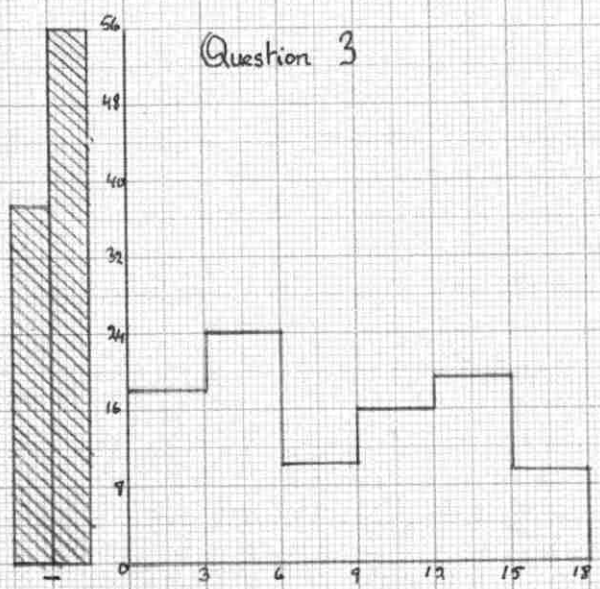
Mid-Year Examination

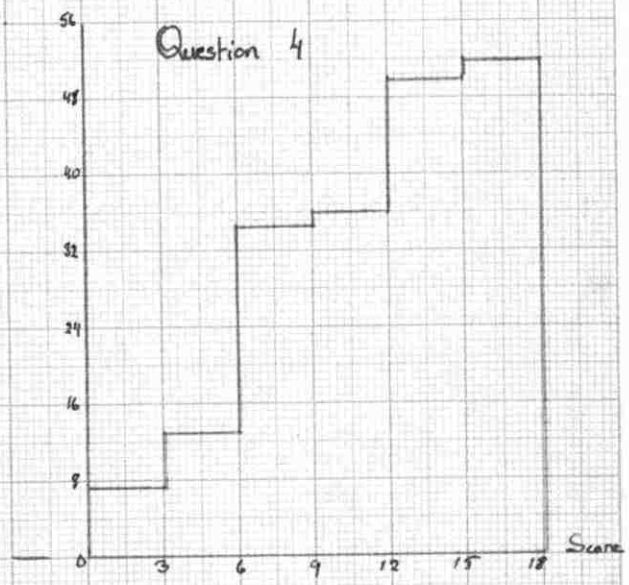


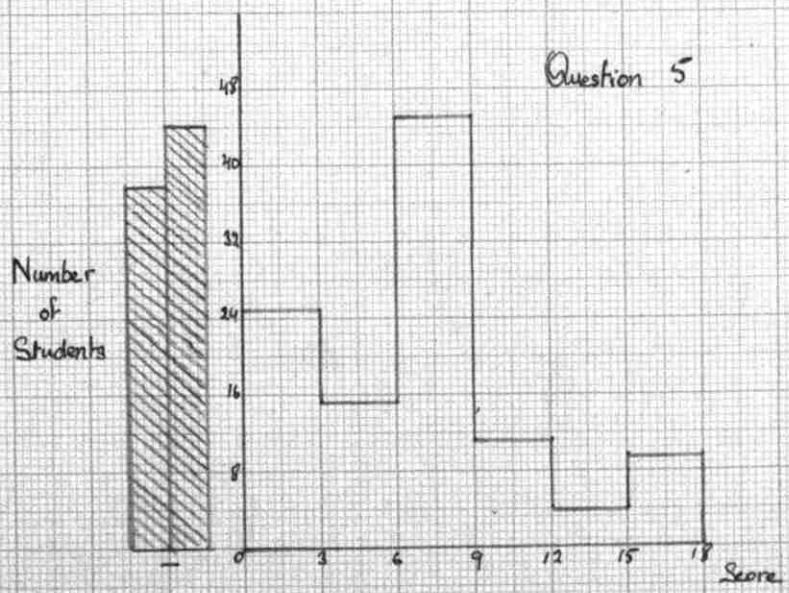
Mid-Year Examination

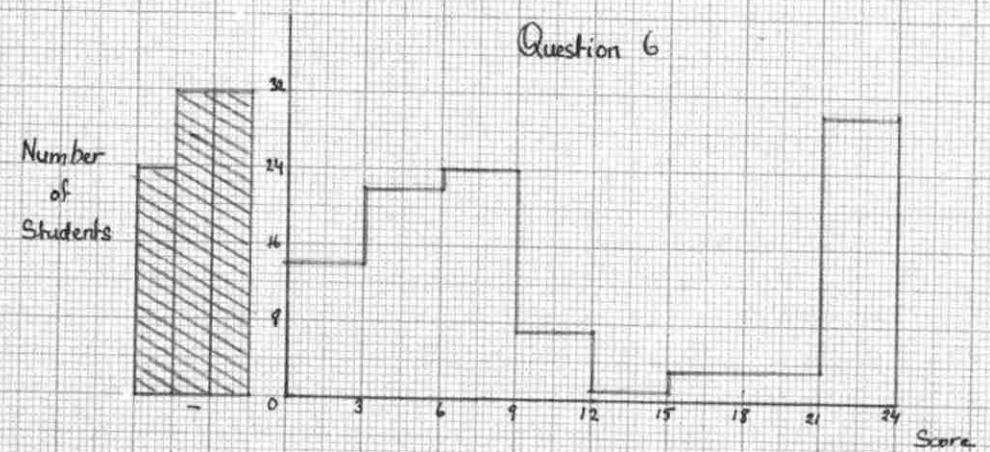


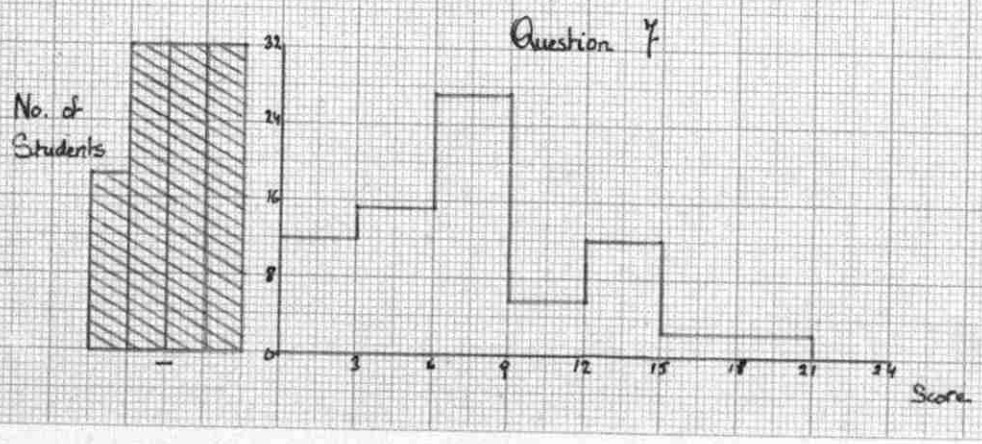






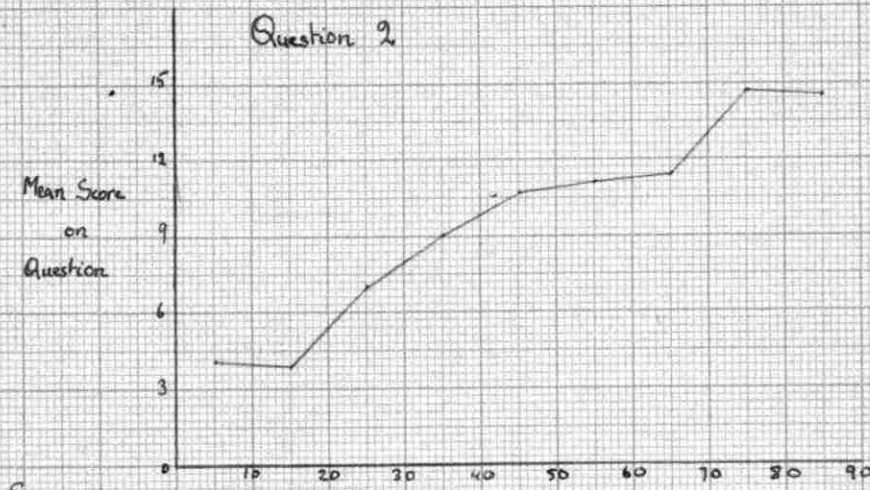


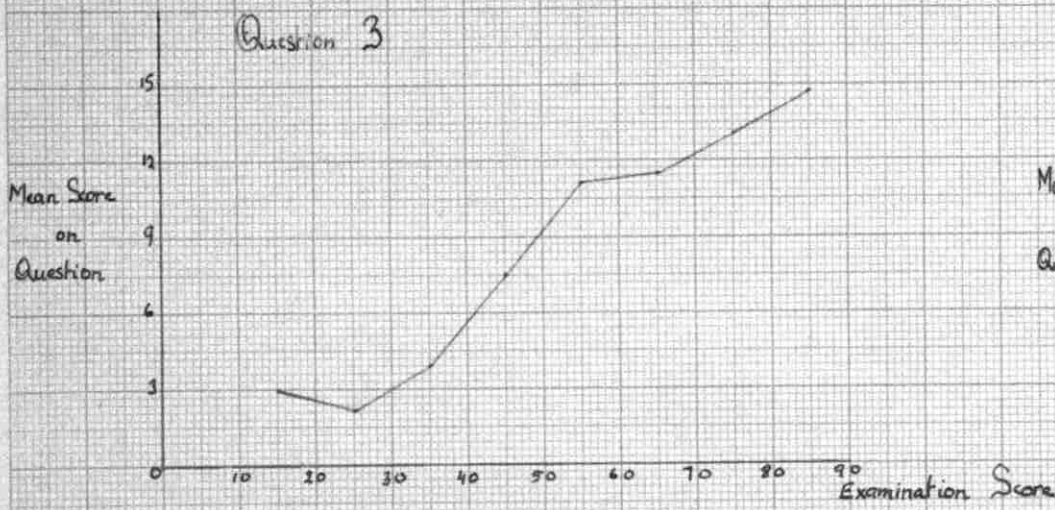

Question 6
Number of Students
32
24
16
8
0
3
6
9
12
15
18
21
24
Score

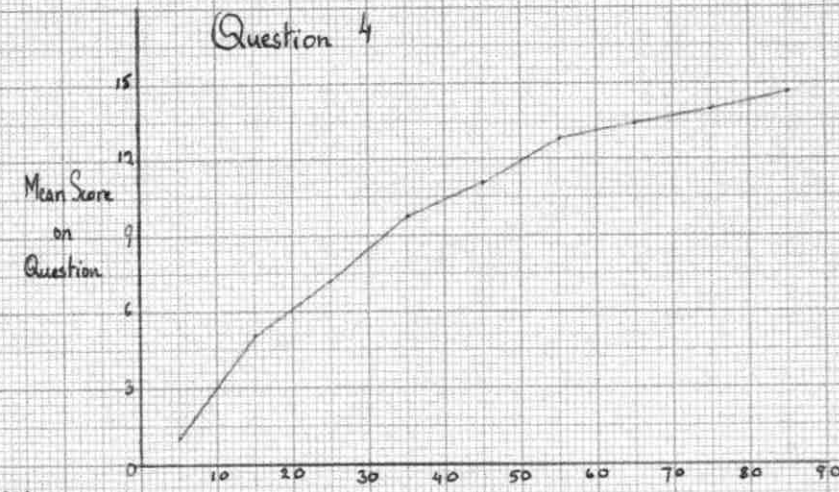

Question 7
No. of Students
32
24
16
8
0
3
6
9
12
15
18
21
24
Score

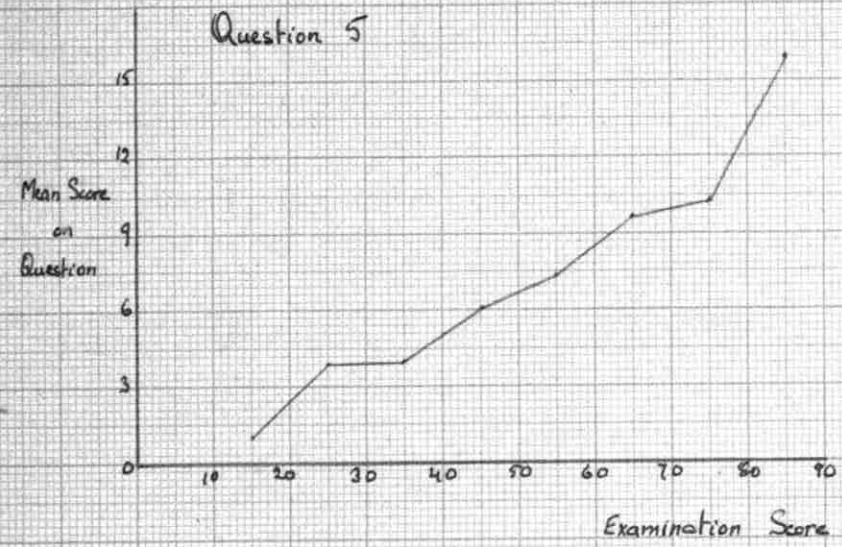
Mid-Year Examination

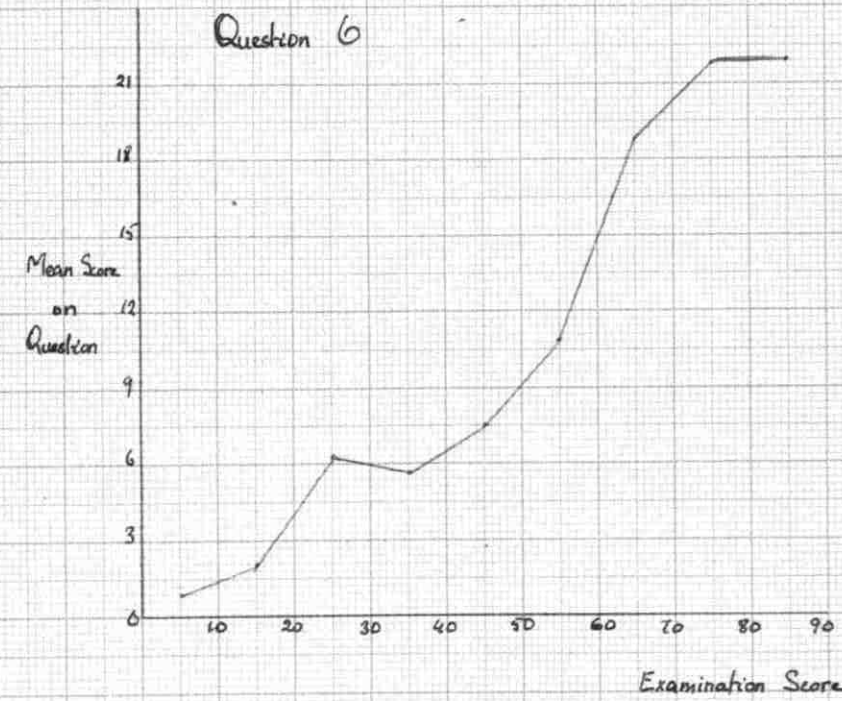
Diagnostic Curves

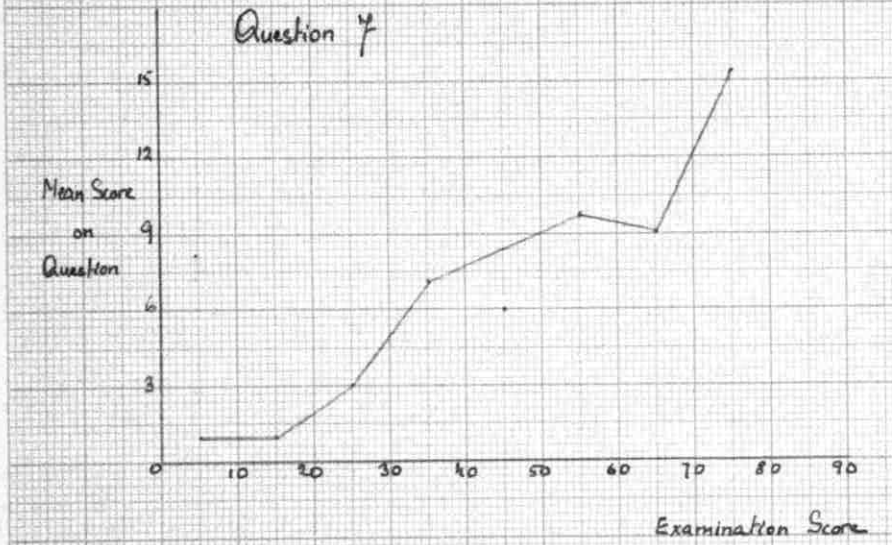







Question 5
Mean Score on Question
0
3
6
9
12
15
10
20
30
40
50
60
70
80
90
Examination Score


Question 6
Mean Score on Question
0
3
6
9
12
15
18
21
10
20
30
40
50
60
70
80
90
Examination Score


Question 7
Mean Score on Question
0
3
6
9
12
15
10
20
30
40
50
60
70
80
90
Examination Score

APPENDIX C

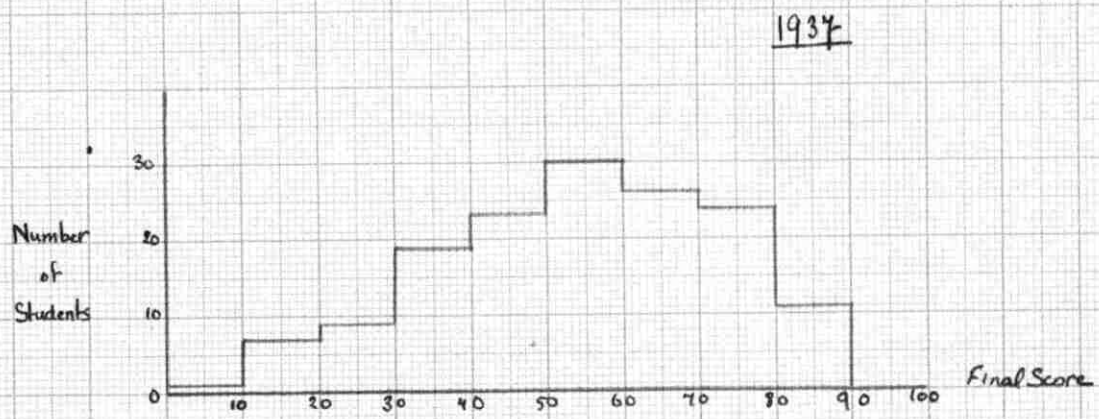
Palestine Matriculation Examinations

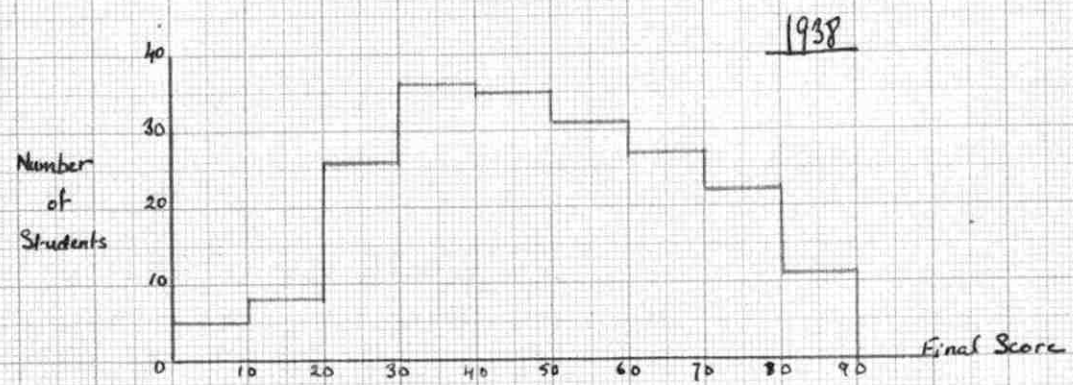
in
Mathematics

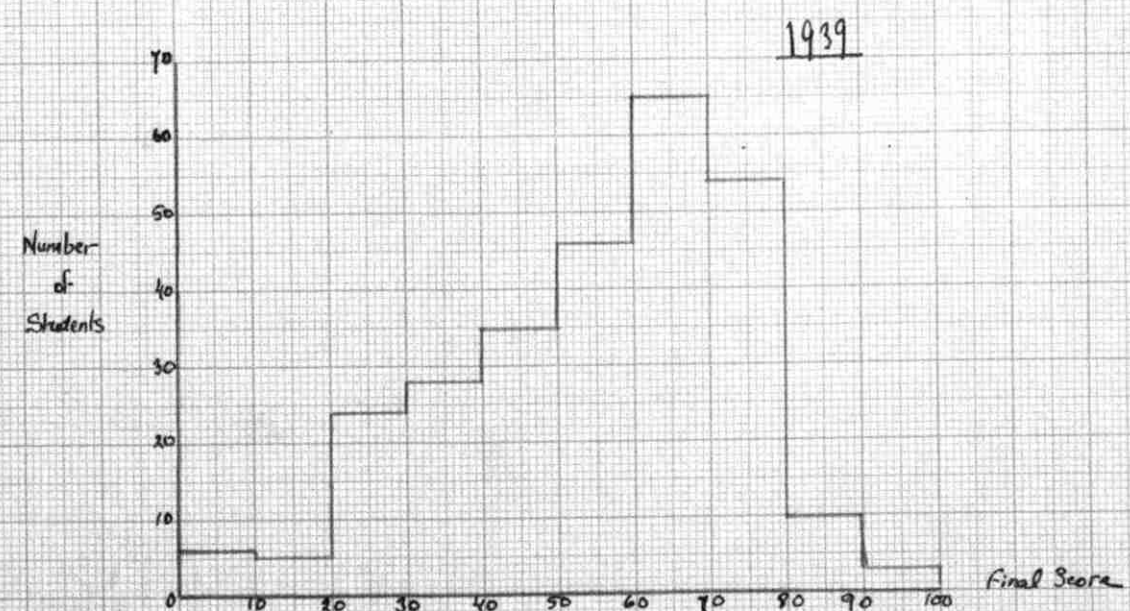
1937-1946

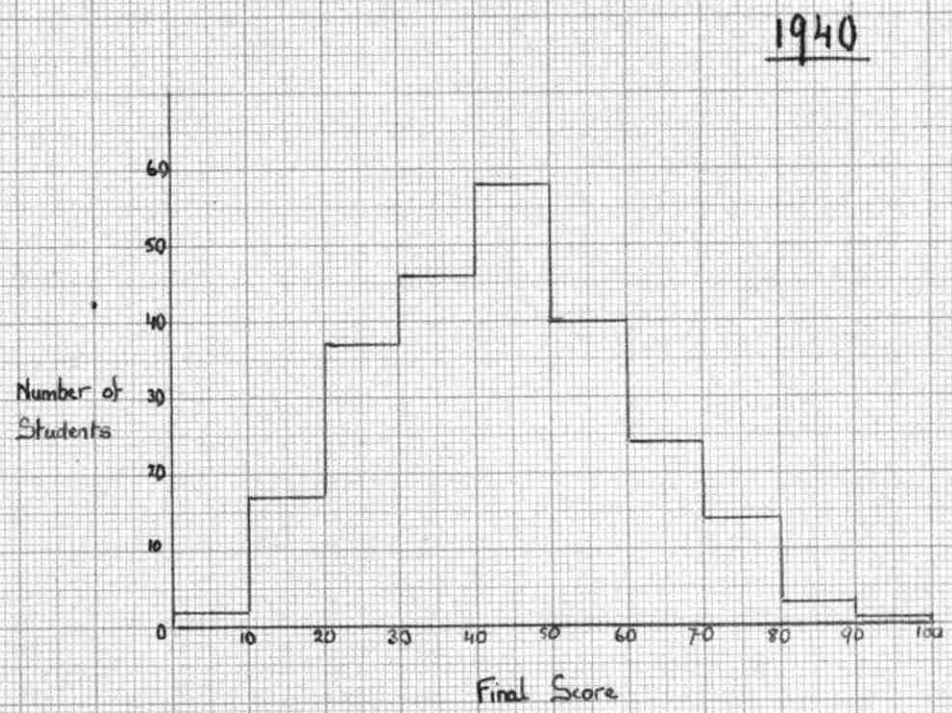
Frequency Distributions of Results

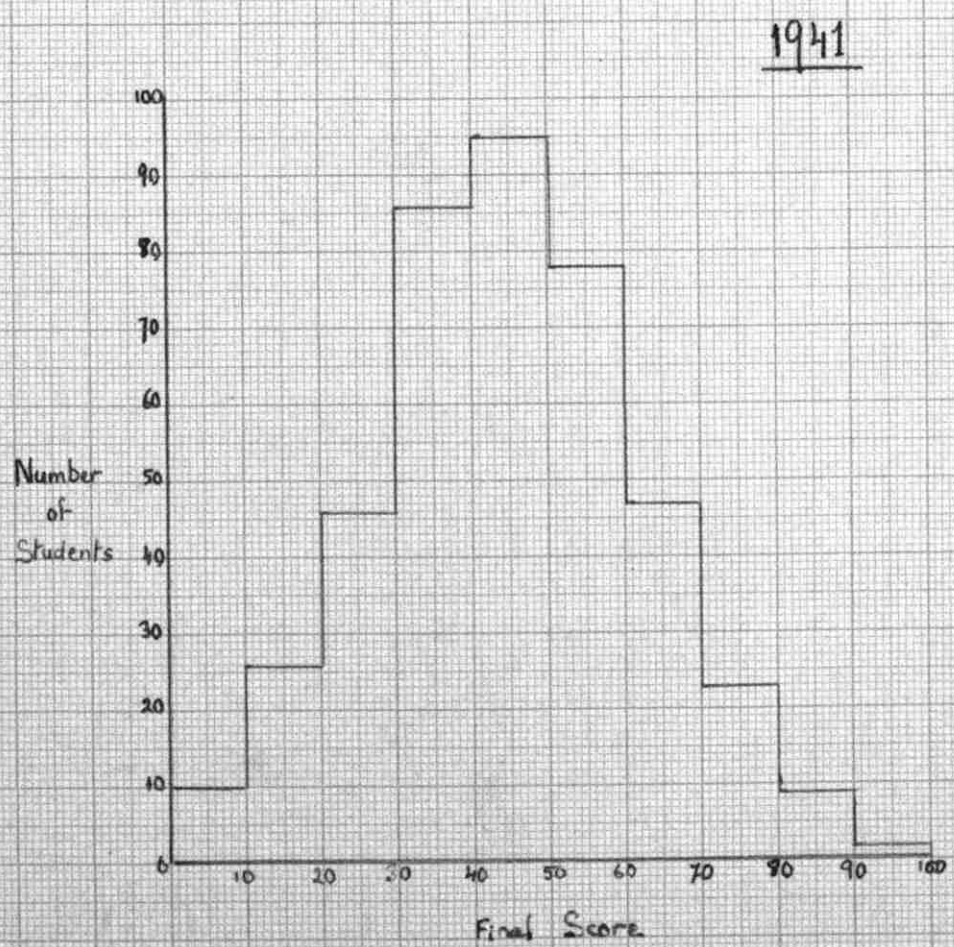
Frequency Distributions of Final Scores.

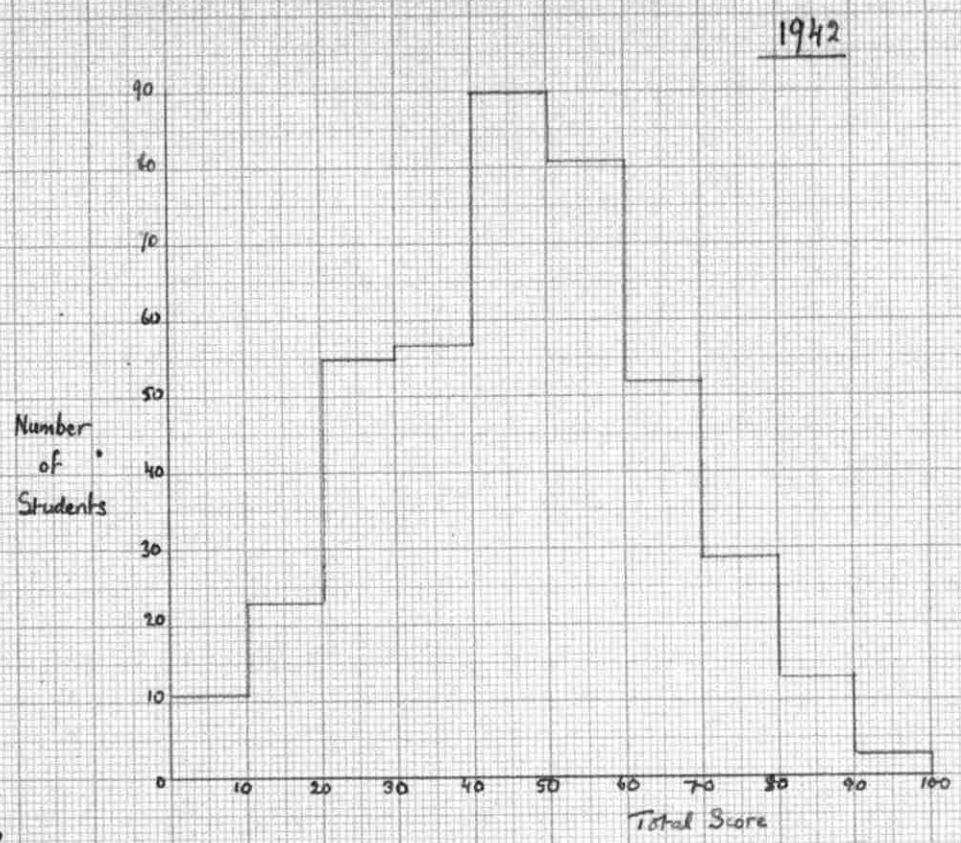


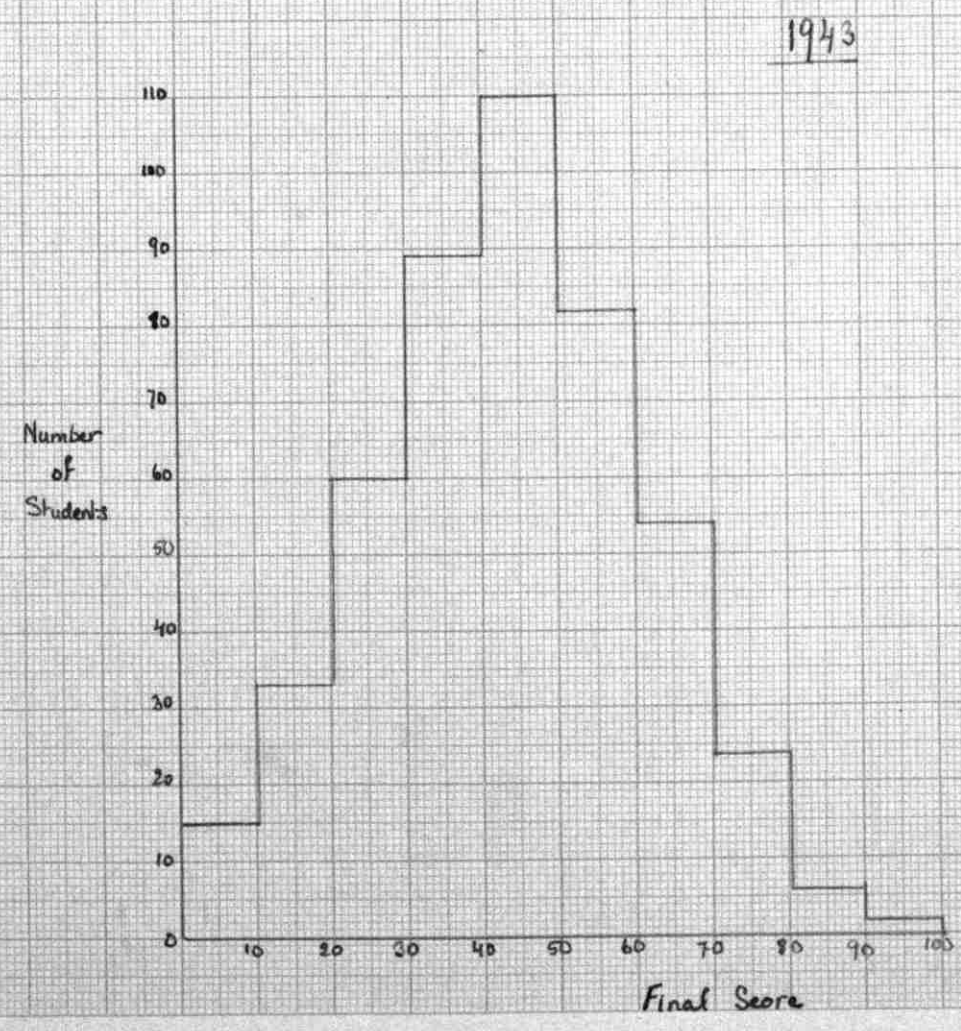




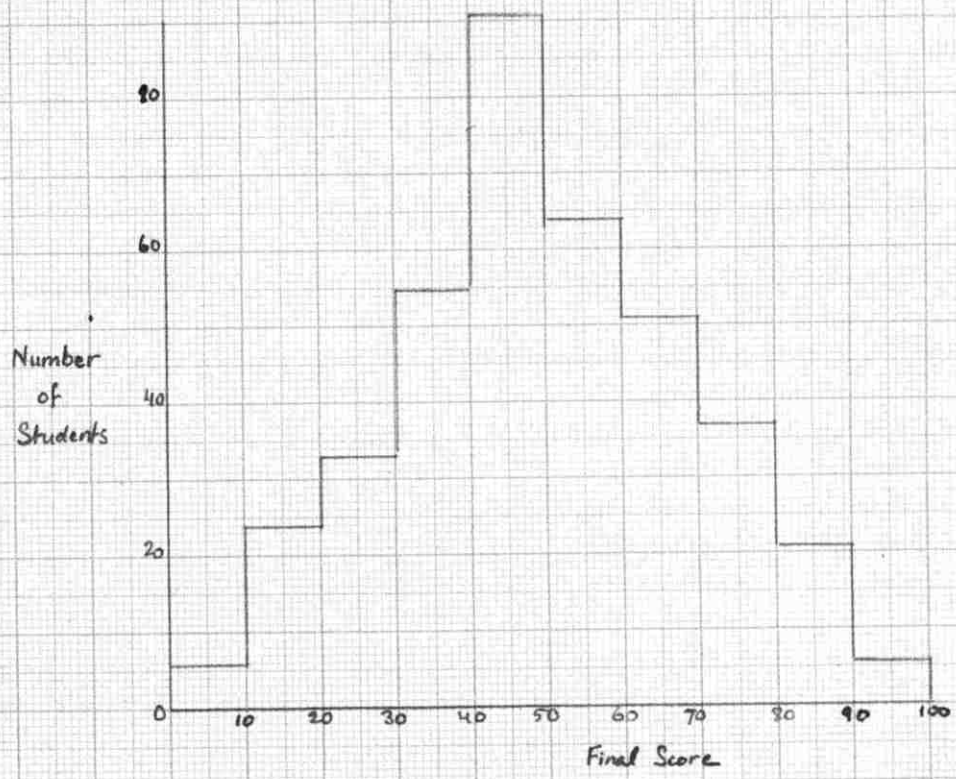

1940
Number of Students
0
10
20
30
40
50
60
10
20
30
40
50
60
70
80
90
100
Final Score


1941
Number of Students
0
10
20
30
40
50
60
70
80
90
100
10
20
30
40
50
60
70
80
90
100
Final Score

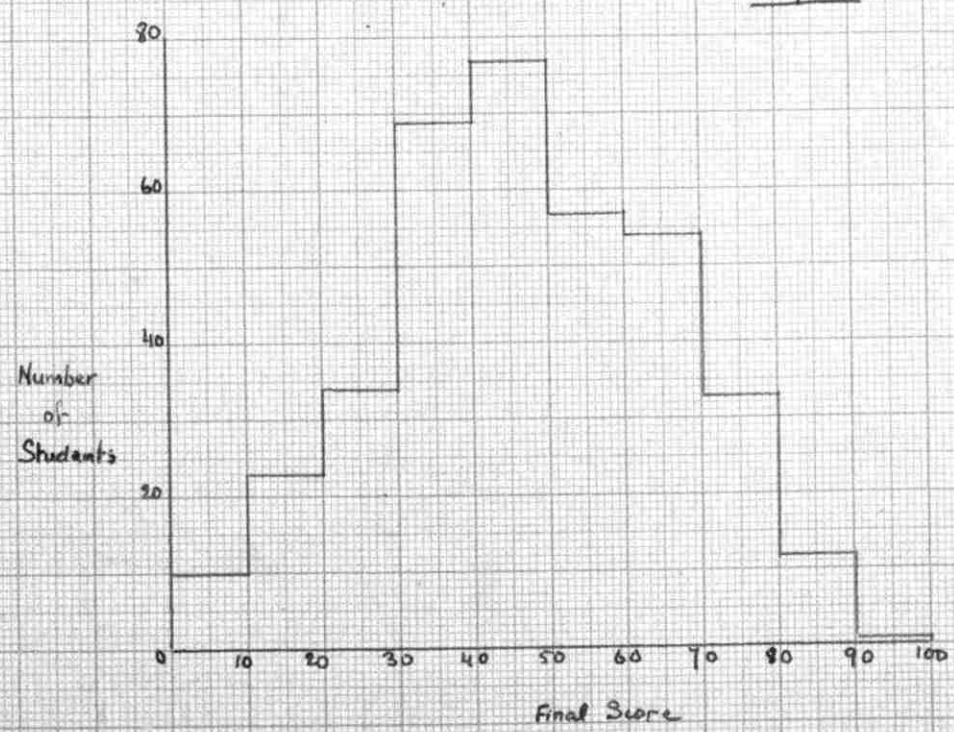

1942
Number of Students
0
10
20
30
40
50
60
70
80
90
10
20
30
40
50
60
70
80
90
100
Total Score

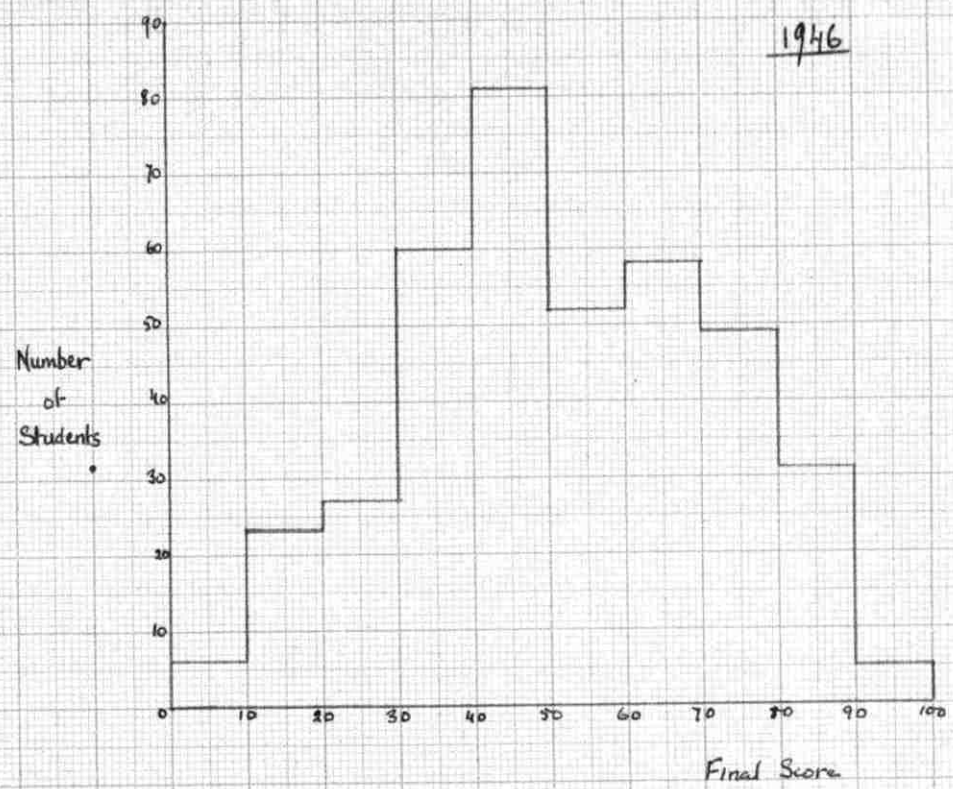

1943
Number of Students
0
10
20
30
40
50
60
70
80
90
100
110
10
20
30
40
50
60
70
80
90
100
Final Score

1944

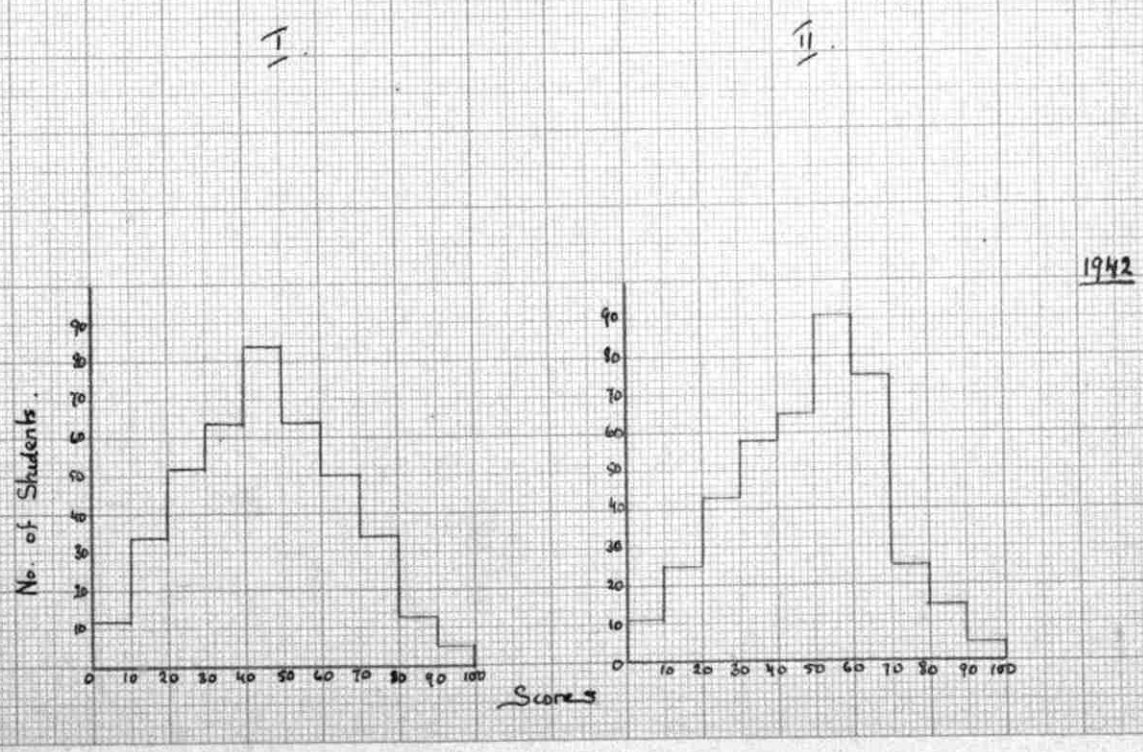

80
60
40
20
0
Number of Students
10
20
30
40
50
60
70
80
90
100
Final Score

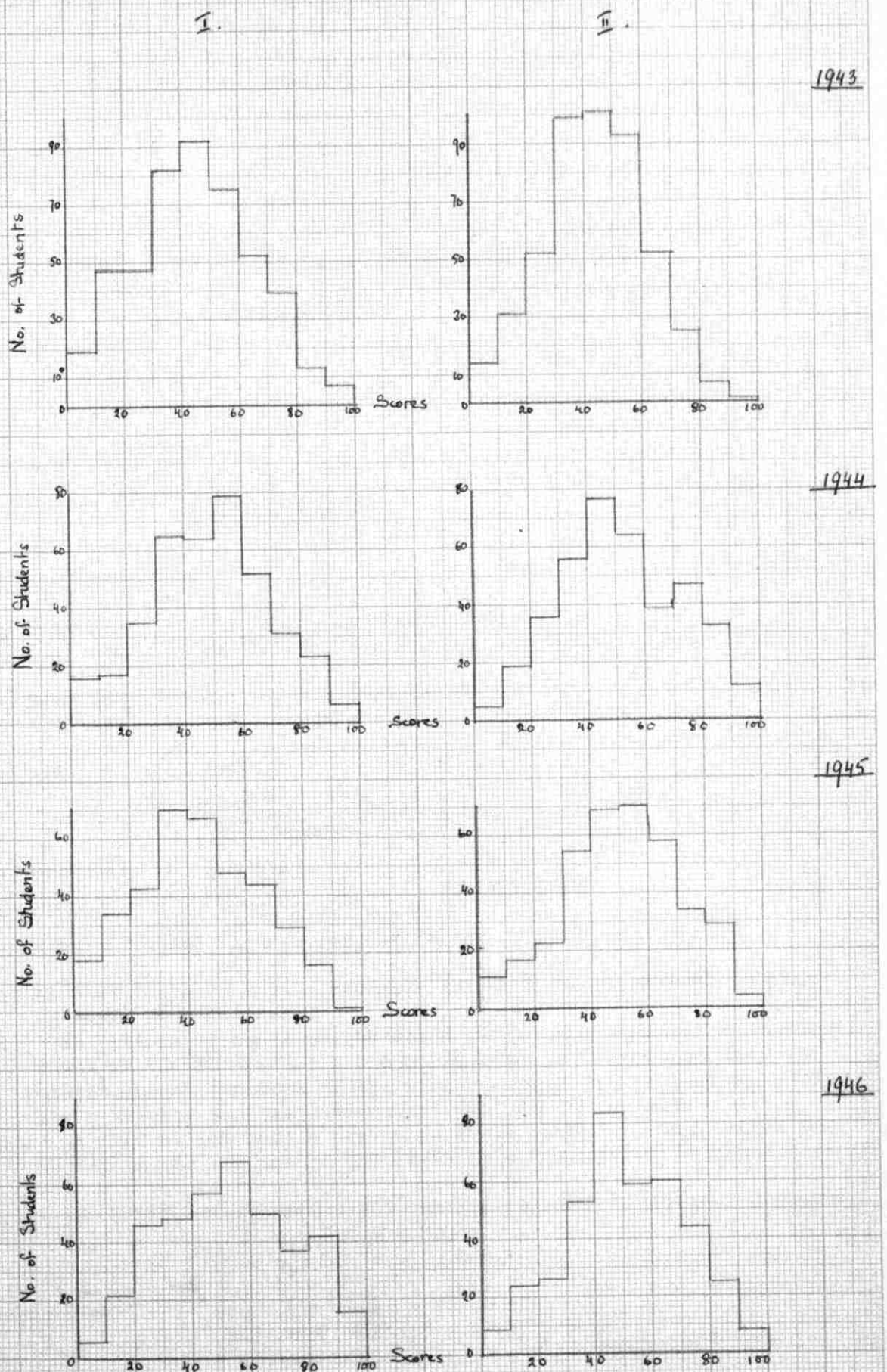
1945


80
60
40
20
0
Number of Students
10
20
30
40
50
60
70
80
90
100
Final Score



Frequency Distributions of Scores on Separate Papers.




I.
II.
1943
1944
1945
1946
No. of Students
Scores
0
20
40
60
80
100

APPENDIX D

A Supplement to

A Brief Study in Elementary Trigonometry Problems

in chapter V.

Question Paper

I. Routine Operations

1. (a) Evaluate log 345.1, log 72.59, log 2247.
   (b) Evaluate log 825.34, log 90.056, log 123.32, antilog 2.2475.

2. (a) Evaluate sin 15, tan 39, cos 78.
   (b) Evaluate any 3 functions (with interpolation)

6. (a) Add

2.95832	0.65513
8.73181 - 10	9.96036 - 10
0.32579	9.57067 - 10

   (b) Subtract

2.92158	3.66305
6.13623 - 10	9.84226 - 10

3. Construct freehand figures for the following problems:

   (i) At what distance does a circular target 6 ft. in diameter subtend an angle of 6 degrees ?
   (ii) Two pillars of equal height stand on opposite sides of a roadway which is 100 ft. wide. A t a point between the pillars the elevations of the tops of the pillars are 71 and 29 degrees. Find their heights and the distance of the point from each.
   (iii) At what height must a person be above earth in order to see an object on the surface of the earth at a distance of 30 miles.?

4. Solve the right triangles(C = 90). Do not include time taken in drawing.
   (i) $c = 25$, $a = 22$;
   (ii) $c = 40$, $A = 15\ 45'$.

5. Factorise:
   (i) $(2x^2 - 4x + 7)^2 - x^2(x + 4)^2$
   (ii) $a^2x^3 - b^2xy^2 - a^2cx^2 + b^2cy^2$
   (iii) $x^4 - 2(b^2 - c^2)x^2 + b^4 - 2b^2c^2 + c^4$.

7. Solve, using the formula $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

(i) $4x^2 + 12x + 13 = 0$
(ii) $(3m + 1)x^2 + 2(m + 1)x + m = 0$
(iii) $49x^2 + 49x + 6 = 0$

8. Prove (i) $\csc A^2 - \cot A^2 = 1$
(ii) $\sin(A + B)/\cos A \cos B = \tan A + \tan B$
(iii) $2 \tan A/(1 + \tan A) = \sin 2A$
(iv) $\frac{\cos A}{1-\tan A} + \frac{\sin A}{1-\cot A} = \sin A + \cos A.$

9. Simplify the expression -
$$(x-3)(2x+5) = x(x+4) + (x+1)(x+3).$$

10. Solve the equation $35x - 24 = 17.$

11. Eliminate x from the equations
$$8x + 4y - 3z = 6$$
$$3x + 9y - 3z = 21$$
$$4x - 5y + 4z = 8.$$

12. Eliminate another unknown from the resulting equations in 11.

II. Recognition Time

1. If a certain number be increased by 16, the result is 7 times the third part of the number. Find the given number.

2. If the radius of the earth is given as 3960 miles, what is the distance around the earth along the equator ?

3. What is the maximum relative error in using 186000 miles/sec. as the speed of light, if it is known that this figure is accurate to the nearest 1000 ?

III. Translation Time

1. The square of the second digit of a number of 3 digits is 11 greater than twice the sum of the first and third digits. The sum of the first and second digits is 4 less than three times the last digit; and, if 99 is added to the number, the digits will be reversed. Find the number.

2. Two circles with their centres on the same diameter of a third circle are tangent to each other externally and to the third internally. The sum of the areas of the inner circles is equal to $\frac{3}{4}$ the area of the outer circle, whose diameter is 16 ins. Find the radii of the smaller circles.

3. A statue 12 ft. high standing on a column subtends, at a point 80 ft. from the base of the column on the same horizontal plane, the same angle as that subtended by a man 6 ft. high, standing at the foot of the column. Find the height of the column.

4. The rate of change of the load in lb./ft. placed on a 20 ft. iron support is 12 times the distance measured in ft. from one end. Find the total load.

IV. Planning Time

1. How far does a point on the equator move in 1 hour because of the rotation of the earth ?

2. Using 3960 miles as the radius of the earth, calculate the distance between two points on the equator if their longitudes differ by an angle of 1 minute.

3. How far from the eye must a coin, 1 inch in diameter, be held in order just to hide the moon, whose angular diameter is 31' 5" ?

4. Prove that the perimeter of a regular polygon of n sides is equal to $2nr \times \tan 180/n$, where r is the radius of the inscribed circle.

5. A regular pyramid has for its base a square whose side is 300 ft. Each face makes an angle of 60 degrees with the base. Find its slant height.

6. From a point on the ground at the foot of a column the angular elevation of the top of a tree is 68 degrees, and at the top of the column 30 ft. from the ground it is 27 degrees. Find the height of the tree and its distance from the column.

Table Showing Individual Results of Timing

Question	Time recorded by every individual							
	1	2	3	4	5	6	7	8
I.								
1 (a) 3 logarithms	1'	1'	2'	2'	1'40"	1'	1'40"	1'15"
(b) with interpolation	2 45	3	4 45	4	3 50	2 30	4 25	4 35
2 (a) 3 Trig. functions	35	55	1 10	1 10	45	45	1 20	55
(b) with interpolation	3 30	4 10	3 10	3	2 40	2 20	4	3 30
3. Constructing figures (average of 3)	35	37	37	46	38	33	50	46
4. Right triangles	3 30	4	4	5	3 20	2 25	2 50	3 10
5. Factoring (i)	1 30	1 45	1 10	1	1 15	1	1 40	45
(ii)	1	1 15	1 30	40	1	45	55	1
(iii)	2	2 20	1 50	1	1 50	45	1 20	1 30
6. (a) adding logs.	1	1	1 45	55	1 55	35	45	35
(b) subtracting logs.	40	1	1 35	45	1	30	40	50
7. 3 quadratics	3 30	4 50	4 30	5 10	5	4 30	1 30	5 30
8. (i) 3 steps	1	1	1 20	1	26	40	45	1
(ii) 4 steps	1 20	1	1 20	1 20	per	1	1	1
(iii) 5 steps	1 30	1 10	1 30	2	step	1 30	1	1 20
9. simplifying(8 terms)	1	1 35	1 30	1 45	45	40	1 30	1 15
10. Linear equation	40	40	30	1	45	20	30	30
11. Unknown in 3 equations	1 30	2 20	2	3 30	1 30	1 40	2 05	2 5
12. Unknown in 2 equations	1 10	1 45	1 30	1 30	1	30	55	55
II.								
1.	30	30			25		15	20
2.	25	35			25		27	20
3.	45	45			35		47	25
III.								
1.	1 45	1 30			1 45		1120	1 15
2.	1 20	45			1		56	55
3.	2	1 55			1 30		1 20	1 10
4.	1	1			55		1	45
IV.								
1. & 2.	30	20			10		30	10
3.	45	40			30		40	25
4.	20	1			35		46	30
5.	15	25			30		8	5
6.	2	1 30			55		1 5	45

Bibliography

A. Educational Statistics.

1. Burt, C. - The Factors of the Mind -
University of London Press, London (1940).
2. Garrett, A.E. - Statistics in Psychology and Education -
Longmans, Green & Co., New York (1926), 3rd. edition (1947).
3. Gregory, C.A. - Fundamentals of Educational Measurement -
D. Appleton & Co., New York (1923).
4. Guilford, J.P. - Psychometric Methods -
McGraw-Hill Book Co., New York (1936).
5. Hartog, P., Rhodes, E.C., & Burt, C. - The Marks of Examinars -
Macmillan & Co. Ltd., London (1936).
6. Holzinger, K.J. - Statistical Methods for Students in Education -
Ginn & Co., U.S.A. (1928).
7. International Institute Examinations Enquiry - Essays on Examinations -
Macmillan & Co., London (1936).
8. Kandel, I.L. - Examinations and Their Substitutes in the United States -
Carnegie Foundation for the Advancement of Learning,
New York (1936).
9. Kent, R.A. & Schruers, E. - "Predictive Value of Four Specified Factors
for Freshman English and Mathematics" -
School and Society, Vol. 27, p. 242, (1928).
10. Leker, C.A. - "Previous Class Cumulative Index as a Guide to Grading" -
Journal of Educational Research, September 1945, pp. 56-61.
11. McCall, W.A.,& Bixler, H.H. - How to Classify Pupils -
Teachers College, Columbia University, New York (1928).
12. Monroe, W.S. - The Constant and Variable Errors of Educational
Measurements -
Bureau of Educational Research, Un. of Illinois (1923).
13. Monroe, W.S. - "Educational Measurement in 1920 and 1945" -
Journal of Educational Research, January 1945, pp. 38-42.
14. Monroe, W.S. & De Voss, J.C. & Kelley, F.J. - Educational Tests and
Measurements -
Houghton Mifflin Co., Massachusets (1924).
15. Odell, C.W. - The Use of Scales for Rating Pupils Answers to Thought
Questions -
Bureau of Educational Research, Un. of Illinois (1929).
16. Paterson, D.G. - Preparation and Use of New-Type Examinations -
World Book Co., New York (1926).
17. Peters, C.C. & Martz, H.B. - "Study of the Validity of the Various
Types of Examinations" -
School and Society, Vol. 27, (1928).
18. Ruch, G.M. - The Improvement of the Written Examination -
Scott, Foresman & Co., U.S.A. (1924).
19. Rugg, H.O. - Statistical Methods Applied to Education -
Houghton Mifflin Co., U.S.A. (1917).

20. Sandon, F. - "Control Charts in Script Assessment in Large Written Examinations" - Journal of the Royal Statistical Society, Vol CVI, Part IV, 1943, pp. 343 - 348.
21. Sims, V.M. - "Reducing the Variability of Examination Marks" - Journal of Educational Research, May 1933, pp. 337 - 349.
22. Sims, V.M. - "Improving the Measuring Abilities of An Essay Examination" - Journal of Educational Research, September 1933, pp. 20 - 31.
23. Spence, R.B., - The Improvement of College Marking Systems - Teachers College, Columbia University, New York (1927).
24. Soderquist, H.O. - "A New Method for Weighting Scores in a True-False Test" Journal of Educational Research, December 1936, pp. 290-2.
25. Sorenson, H. - Statistics for Students of Psychology and Education - McGraw-Hill Co. Inc., New York (1936).
26. Stevason, C.C. - "Simplifying the School Marking Process" - Journal of Educational Research, April 1945, p. 64 - .
27. Strang, R. - "Correlation of Reading and Intelligence Test Scores" - Journal of Educational ~~of~~ Research, Feb. 1945, pp. 440-45.
28. Traxler, A.E. - "Problems Arising out of the Attempt to Apply Improved Measurement Techniques to Education & Guidance" - Journal of Educational Research, September 1943, pp. 14-18.
29. Valentine, C.W. - The Reliability of Examinations - Un. of London Press, London (1923).
30. Vernon, P.E. - The Measurement of Abilities - Un. of London Press, London (1940).
31. Wood, B.D. - Measurement in Higher Education - World Book Co., New York (1923).

B. Statistics Reference Books.

1. Aitken, A.C. - Statistical Mathematics - Oliver & Boyd, Edinburgh (1939), 3rd. edition (1944).
2. Baten, W.D. - Elementary Mathematical Statistics - John Wiley & Sons, New York (1938).
3. Davenport & Ekas - Statistical Methods in Biology, Medicine & Psychology - John Wiley & Sons, New York (1936).
4. Davies, G.R. & Crowder, W.F. - Methods of Statistical Analysis in the Social Sciences - John Wiley & Sons, New York (1933).
5. Ezekiel, M. - Methods of Correlation Analysis - John Wiley & Sons, New York (1930).
6. Fisher, R.A. - The Design of Experiments - Oliver & Boyd Ltd., Edinburgh (1942).
7. Holzinger, K.J. & Harman, H.H. - Factor Analysis - Un. of Chicago Press, Illinois (1941).
8. Kendall, M.G. - The Advanced Theory of Statistics - Charles Griffin & Co. Ltd., London (1945).
9. Peters & Van Voorhis, W.R., - Statistical Procedures and Their Mathematical Bases - McGraw-Hill Book Co., New York (1940).

10. Snedecor, G.W. - Calculation and Interpretation of Analysis of Variance and Covariance - Collegiate Press Inc., Ames, Iowa (1934).
11. Tippett, L.H.C. - The Methods of Statistics - Williams & Norgate, London (1931), edition (1941).
12. Weatherburn, C.E. - A First Course in Mathematical Statistics - Cambridge University Press (1946).
13. Yule, G.U. - An Introduction to the Theory of Statistics - Charles Griffin & Co., London (1932).

C. Statistical Tables

Fisher, R.A. & Yates, F. - Statistical Tables for Biological, Agricultural and Medical Research - Oliver & Boyd, Edinburgh (1938).

ERRATA AND CORRIGENDA

Page	Error	Correction
2 l. 15	The normal distribution curve	The reduced normal distribution curve
11 para 3	The marking of an interval on the percentage marks scale .... above that particular interval	The marking of an interval on the percentage marks scale is proportional to the area under the curve corresponding to that particular interval. Foa a scale ranging between 66 and 99, 17 marks lie below and 17 marks above the mean, 82.5. The whole range of values, in terms of σ, on each side of the mean is divided into $16\frac{1}{2}$ equal intervals and the sections of area above each of these divisions read from normal curve tables, the spaces are then marked on the scale in proportion to the consecutive sections of area.
para 4	Since this method causes scores to accumulate about the centre, ... and leave the separate items with unaltered scores	By the use of this scale all sets of marks for a class can be converted to sets normally distributed with the same mean and variance. Considering the marks in these sets as independent variates, their means should also be normally distributed about the common mean but with 1/n of the original variance, where n is the number of sets (cf. Weatherburn p. 58). If, then, the individual scores of n separate items or tests are first referred to the normal curve by Stevason's scale, and then added, the final averages will be crowded about the centre, particularly where n is large. In this case, it may be better to convert the totals by this method and leave the separate items with unaltered scores.
16 para 4	but also implies an equal degree of non-correlation.	but also implies absence of correlation to the same extent; since the coefficient of alienation, $k = \sqrt{1 - r^2} = \sqrt{1 - .49} = \sqrt{.51} = .7$ (app), is taken to measure the degree of lack of correlation just as r measures its presence.
18 line 37	$\alpha = m/m-2\Sigma x_j^2 -2/m-2\Sigma x_j(y_j+x_j+z_j+..) -1/(m-1)(m-2)[\Sigma x_j^2+\Sigma y_j^2..] +1/(m-1)(m-2)\Sigma(x_j+y_j..)^2$	$\alpha = [m/(m-2)]\Sigma x_j^2 - [2/(m-2)][\Sigma x_j(y_j + x_j + z_j + ...)] - [1/(m-1)(m-2)](\Sigma x_j^2 + \Sigma y_j^2 + ...) + [1/(m-1)(m-2)]\Sigma(x_j + y_j + ...)^2$
19 line 9	a perfect of a zero score	a perfect or a zero score

Page		Error	Correction
21	line 5	is taken at -5	is taken at -5σ
25	(3)	where w is the weight assigned to x	where w_i is the weight assigned to x_i
26	para 4	the upper and lower quartiles, q and q	the upper and lower quartiles, q_1 and q_3
27	note (3)	P. E. = 0.6745	P.E. = 0.6745σ
	" (4)	The third moment is divided by s	The third moment is divided by s^3
28	(12)	$y = \frac{\sum_i x_i' y_i'}{s_x^2} x$	$y' = \frac{\sum_i x_i' y_i'}{s_x^2} x'$
	last line	$\sum_i (y_i - bx_i)^2$	$\sum f_i (y_i - bx_i)^2$
29	(17)	where y is the value of y in the ith row and jth column,	where y_{ij} is the value of y in the ith row and jth column
	last line	The formula follows directly from the product moment coefficient formula.	The formula follows from the product moment coefficient formula if we regard ranks as variate values, and then obtain the mean and variance of each and the covariance in terms of n, the number of variates. These, when substituted in the formula for the product moment coefficient, give (18).
30	para 2	The standard error, S.E., is defined as the standard deviation of residuals..... ..... is the S.E. of the mean.	The standard error, S.E., of a statistic is defined as the standard deviation of the sampling distribution function of that statistic, the samples being drawn at random from the same parent population. If the mean of a population is estimated from the sample mean m, the standard error involved is the standard deviation of m, since, as the number of samples increases, the grand mean $\bar{m}$ tends to approach the true population mean in value.
31	line 1	$d(\frac{1}{2}\chi^2)$	$d(\chi^2)$
	" 3	x_i's are deviations of n independent variates	x_i's are deviations of n observed frequencies
	" 6	e_i is the expected value of x_i	e_i is the expected value of the frequency in the ith class
	" 9	a normal distribution	a normal distribution for χ
	" 19	If χ^2 is significant, the hypothesis of a non-fit .... and the fit is poor	If the value of χ^2 obtained is significantly high (with a very low probability of occurence) it implies that the discrepancies are not merely sampling effects. If the probability of its being exceeded in the same distribution is high, χ^2 is insignificant, and the values show a "good fit".

Page		Error	Correction
31	line 23	the ratio $(m-\mu)/SE_m$ is normally distributed	the ratio $(m-\mu)/SE_m$ is approximately normally distributed
35	table 3	(4th column heading) d	d^2
39	para 3 line 9	For, if we estimate achievement scores from intelligence scores,	For, assuming a linear regression of achievement on intelligence scores, if we estimate the former from the latter,
40	line 13	= (1/496)( )	= (1/512)( )
44	line 9	and n is the total number of answers	and n is the total number of answers, i.e. $n = r + w$
	" 18	$E(s) = \frac{2n-n}{n-1}$	$E(s) = \frac{2n-n}{n}$
	" 19	(add)	And if a student, unwilling to guess, leaves unanswered some of the statements, the total credit he then obtains on the whole question is proportional to the number of attempted answers; i.e., if the number of expected answers is N, he receives only na/N of the total credit, n and s being defined as above.
46	line 8	adjusted score = original score + (100 - original score)	adjusted score = original score + ¼(100 - original score)
47	line 4	Each point on the graduated scale .... erected at that point.	The distance between two points on the graduated scale denotes the percentage of the total area below the curve and between the two corresponding abscissae.
52	line 9	91.4 - (8.7)	$91.4 - (8.7)^2$
58	line 15	as compared with .002 in accurate calculation	(omit)
61	para 3	If r is calculated..... ..... all this accuracy.	(omit)
67	line 20	At the 5% level = .032 At the 1% level = .052	At the 5% level $\varepsilon^2 = .032$ At the 1% level $\varepsilon^2 = .052$
72	line 2	Constructive is just ready to begin	Constructive research is just ready to begin
74	line 4	$(1/\sqrt{n}$, med)	$(1/\sqrt{n_i}$, $med_i)$
77	line 6	$m - \mu/\sigma_m$	$(m - \mu)/\sigma_m$
	" 9	$\frac{m - \mu\sqrt{n}}{s}$	$\frac{(m - \mu)\sqrt{n}}{s}$

Page	Error	Correction

85 (Table of calculations)

x	$z=\frac{x-\mu}{\sigma}$	$\frac{1}{2}$ erf z	$\frac{1}{2}\Delta$erf z	Normal Scale	Arbitrary Scale
0	-3.743	-.5000			
			.0459	5	5
305	-1.686	-.4541			
			.2042	20	20
455	-0.674	-.2499			
			.4998	50	50
655	0.674	.2499			
			.2286	23	20
855	2.024	.4785			
			.0215	2	5
100	3.002	.5000			

86- last line 87 For, assuming normal distributions of marks given on papers I and II,..... .... which is not a normal function.

A normal frequency distribution of a sum of variates results when the conditions are satisfied that the elements that add up to the sum be <u>both</u> normally distributed and be independent of each other. For, if the distribution functions of papers I and II were -

$$F(x) = 1/\sqrt{2\pi}\,\sigma_x \int_{-\infty}^{x} \exp\left[-\tfrac{1}{2}\left(\frac{x-m_x}{\sigma_x}\right)^2\right] dx$$

$$F(y) = 1/\sqrt{2\pi}\,\sigma_y \int_{-\infty}^{y} \exp\left[-\tfrac{1}{2}\left(\frac{y-m_y}{\sigma_y}\right)^2\right] dy\ ,$$

then the distribution function of (x+y) would be

$$F(z) = 1/\sqrt{2\pi(\sigma_x^2+\sigma_y^2)} \int_{-\infty}^{z} \exp\left[-\tfrac{1}{2}\frac{(z-m_x-m_y)^2}{\sigma_x^2+\sigma_y^2}\right] dz\ .$$

(cf. Weatherburn, p58)

In the absence of the condition of independence, the probability function of the sum of two elements is not the simple product of the probability functions of the elements. Also, when one of them is not normally distributed, a normally distributed sum does not result.

In this case, the results of papers I and II are expected to be highly correlated; and, at the same time, there is the possibility that only one of the papers has normally distributed results.

87 table 1 (columns I & II)

$\chi^2 = 2.43$, $P > .50$ $\chi^2 = 6.23$, $P > .10$
$\chi^2 = 41.82$, $P < .01$ $\chi^2 = 1.89$, $P > .50$
$\chi^2 = 25.03$, $P < .01$ $\chi^2 = 1.4$, $P > .50$

88 para 5 (calculations)

$m_A = 51.2$, $m_H = 35.03$, $m_A - m_H = 11.4$;
$s_A = 294.1$, $s_H = 281.1$,

$$\sigma_{M_A-M_H} = \sqrt{\frac{334.95}{124} + \frac{289.2}{155}} = 2.137.$$

$$\frac{m_A - m_H}{\sigma_{M_A-M_H}} = \frac{11.4}{2.137} = \underline{5.335.}$$

Page		Error	Correction
89	para 4	$m_A = 0.67$ $m_E = -.13$ $M = -2.99$ $m_H = -9.47$	$m_A = 0.67$ $m_E = -.13$ $M = -.02$ $m_H = -.47$
		$5x2 + 4(9-10) + \ldots\ldots$ $\ldots\ldots\ldots 1(78-86)/N$	$\left[5x2 + 4(9-10) + 3(23-26)\ldots\ldots + 1(78-86)\right]^2/N$
		$=(10 - 4 - 9 + 2 - 8)/422$	$= (10 - 4 - 9 + 2 - 8)^2/422$
		$= 1331 - .2 = \underline{1330.8}$	$= 1331 - .19 = \underline{1330.81}$
		$\sum(m_j - M)^2 = 3.66 + 2.86 + 6.48$ $= 63.56$	$\sum_j \frac{(\sum_i x_{ij})^2}{f_{x_j}} - .19 = (83/124 + 19/143 + 73/155) - .19$ $= 55.56 + 2.64 + 34.38 - .19$ $= 92.58 - .19$ $= 92.39$
		$\sum_i\sum_j(x_{ij} - \bar{x}_{ij})^2 = 13308 - 6356$ $= \underline{1266.36}$	$\sum_i\sum_j(x_{ij} - m_j)^2 = 1330.81 - 92.39$ $= \underline{1238.42}$

table 2

Sum of Squares	Mean S.	Sum of Squares	Mean Squares
1330.8	3.16	1330.81	3.16
63.56	31.78	92.39	46.19
1266.36	3.022	1238.42	2.96

90	line 1	$F = 31.78/3.022 = 10.52$	$F = 46.16/2.96 = 15.61.$
	" 5	(10.52)	(15.61)
	" 16	$(A, E, H, i=1,2,\ldots)$	$(A_i, E_i, H_i, i = 1, 2, \ldots)$
96	line 5	promise better success is larger	promise better success if larger
114	3 l.4	$(y - a - bx)^2$	$\sum(y - a - bx)^2$
115	5	Correction in r for Interval Grouping	(omit)