UNIVERSAL
LIBRARY
OU_150450
UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 371.132/H33M Accession No. 25761

Author Hartog, P. & Rhodes, E.C.

Title Marks of Examiners

This book should be returned on or before the date last marked below.

THE MARKS OF EXAMINERS

INTERNATIONAL INSTITUTE EXAMINATIONS ENQUIRY

Members of the Committee :

SIR MICHAEL SADLER, K.C.S.I., C.B., LL.D. (*Chairman*)
Sometime Master of University College, Oxford, and Vice-Chancellor of the University of Leeds.

P. B. BALLARD, M.A., D.Lit.
Sometime Inspector in the Education Department of the London County Council.

C. DELISLE BURNS, M.A., D.Lit.
Stevenson Lecturer in Citizenship in the University of Glasgow.

CYRIL BURT, M.A., D.Sc.
Professor of Psychology in the University of London.

H. R. HAMLEY, M.Sc., Ph.D.
Professor of Education in the University of London.

SIR PHILIP HARTOG, K.B.E., C.I.E., LL.D. (*Director*)
Sometime Vice-Chancellor of the University of Dacca and Chairman of the Auxiliary Committee on Education of the Indian Statutory Commission.

SIR PERCY NUNN, D.Sc., Litt.D.
Professor of Education in the University of London.

C. SPEARMAN, LL.D., F.R.S.
Emeritus Professor of Psychology in the University of London.

GODFREY H. THOMSON, D.Sc.
Professor of Education in the University of Edinburgh.

F. CLARKE, M.A.
Professor-elect of Education in the University of London.

Other publications of the International Institute Examinations Enquiry Committee :—

An English Bibliography of Examinations (1900-1932) by Mary C. Champneys, with a Foreword by Sir Michael Sadler and Sir Philip Hartog (pp. xxiv, 141), 1934. Price 5/-.

An Examination of Examinations, by Sir Philip Hartog and Dr. E. C. Rhodes (Second Edition (Third Impression), pp. 81), 1936. Price 1/-. (First published 1935.)

Essays on Examinations, by Sir Michael Sadler, A. Abbott, P. B. Ballard, Cyril Burt, C. Delisle Burns, Sir Philip Hartog, C. Spearman and S. D. Stirk (pp. xii, 168). 1936. Price 5/-.

To be published shortly :—

A Conspectus of Examinations conducted in Great Britain and Northern Ireland.

INTERNATIONAL INSTITUTE EXAMINATIONS ENQUIRY

THE MARKS OF EXAMINERS

Being a Comparison of Marks allotted to Examination Scripts by Independent Examiners and Boards of Examiners, together with a Section on a Viva Voce Examination

BY

SIR PHILIP HARTOG, K.B.E., C.I.E.

AND

E. C. RHODES, D.Sc.

READER IN STATISTICS IN THE UNIVERSITY OF LONDON

WITH A MEMORANDUM BY

CYRIL BURT, M.A., D.Sc.

PROFESSOR OF PSYCHOLOGY IN THE UNIVERSITY OF LONDON

MACMILLAN AND CO., LIMITED
ST. MARTIN'S STREET, LONDON
1936

TABLE OF CONTENTS

PAGE

PREFACE vii

PART I

BY

P. J. HARTOG AND E. C. RHODES

CHAPTER

I Marking of School Certificate History Scripts (Two Investigations) 1

II Marking of School Certificate Latin Scripts . . . 17

III Marking of School Certificate French Scripts . . . 34

IV Marking of School Certificate Chemistry Scripts . . 51

V Marking of School Certificate English Scripts . . . 64

VI Special Place Examination (I): Marking of Arithmetic and English Scripts

Section I. Introductory 68

Section II. The Combined Results of the Examinations in Arithmetic and English 70

Section III. Arithmetic 76

Section IV. English Essay Marks 83

Section V. English, Part B 93

Section VI. English, Part B—Detailed Examination of the Marks awarded for Parts of a Question . . 101

Appendix I to Chapter VI 112

Appendix II to Chapter VI 114

VII Special Place Examination (II): Marking of English Essay 117

Appendix to Chapter VII 138

VIII College Entrance Scholarship Examination: Marking of English Essay 142

IX Marking of University Mathematical Honours Scripts . . 148

X Marking of University History Honours Scripts . . . 152

XI A Viva Voce (Interview) Examination 168

Appendix to Chapter XI 177

PART II

BY

E. C. RHODES

CHAPTER		PAGE
XII	On Differences of Standard and Random Variations	179
	Notes to Chapter XII	
	Note I—Connection between Correlation Coefficients and the Size of the Random Element in Marking	195
	Note II—On the Assumption that there is no Connection between Random Variations of Different Examiners	197
	Note III—On the Spreading of Ideal Marks	198
XIII	Results of Applying the Method of Analysis to the Data of the Investigations	
	Section 1. School Certificate History	199
	Section 2. School Certificate French	204
	Section 3. School Certificate Chemistry	210
	Section 4. School Certificate English	216
	Section 5. School Certificate Latin	223
	Section 6. English Scholarship Essay	228
	Section 7. History Honours	231
	Section 8. Mathematical Honours	234
	Section 9. Essay Scripts at the Special Place Examination (II)	239
	Section 10. Special Place Examination (I)	240
	Section 11. Summary of the Results	242

MEMORANDA

MEMORANDUM		
I	The Analysis of Examination Marks. By Cyril Burt	
	Section I. Introduction	245
	Section II. Preliminary Assumptions	246
	Section III. Adjusting for Differences of Scale: the Average and the Standard Deviations	266
	Section IV. Measuring the Individual Examiner's Accuracy: the Coefficient of Correlation	270
	Section V. To Determine the Correlation between the Marks of a Given Examiner and the Hypothetical True Marks: the "H.G.F." or "Saturation Coefficient"	280
	Section VI. To Determine the Hypothetical "True" Mark for a Given Candidate: the Weighted Average	297
	Section VII. Specific Factors	304
	Section VIII. Summary	309
	Note to Memorandum I	312
II	A Second Approximation for the Determination of Ideal Marks and Random Variations. By E. C. Rhodes	315
III	On Certain Points of Difficulty in Connection with School Certificate Examinations. By P. J. Hartog	325
IV	A Reply to Some Criticisms of *An Examination of Examinations*. By P. J. Hartog and E. C. Rhodes	337

ERRATA

Graphs :

PAGE

21 School Certificate Latin—Group I. (1) On the vertical line representing Candidate No. 11, the letter E on the 40 mark horizontal line should be replaced by D. (2) On the vertical line representing Candidate No. 10, the letter D should be inserted on the 37 mark horizontal line.

22 School Certificate Latin—Group II. (3) In respect of Candidate No. 8, M and N should be inserted on the 48 mark horizontal line instead of on the 49 mark as shown.

62 School Certificate Chemistry. (4) In respect of Qn. 1, E should be inserted on the same level as D. (5) In respect of Qn. 2, E should be inserted below the 40 mark horizontal line (39·6). (6) In respect of Qn. 4, the positions of A and E should be interchanged. (7) In respect of Qn. 6, the positions of K and L should be interchanged.

90 Special Place Examination (I) : English Essay. (8) In the vertical line corresponding to Craftsmanship, the positions of A and J should be interchanged.

174 Viva Voce Examination. (9) In respect of Candidate No. 3, the letter D should be inserted on the 150 mark horizontal line. (10) In respect of Candidate 6, the letter B should be inserted on the 150 mark horizontal line instead of on the 140 mark line.

ADDENDUM

Add on p. 333, line 2 from top, after the words "May, 1936," the following :—

Some particulars of the method used by the Cambridge Local Examinations Syndicate are given in an article on "The Reliability of School Certificate Results," by Mr. J. O. Roach, Assistant Secretary of the Syndicate, in *Oversea Education* for April, 1936.

PREFACE

(i) No element in the structure of our national education occupies at the present moment more public attention than our system of examinations. It guards the gates that lead from elementary education to intermediate and secondary education, from secondary education to the Universities, the professions, and many business careers, from the elementary and middle stages of professional education to professional life.

(ii) Quite apart from the safeguards imposed by Acts of Parliament and Government authorities, a whole congeries of examinations has sprung up in the last century, created by private and public bodies.[1] Examinations have become a familiar topic in our newspapers and in our homes. The examination system has grown to be an important element, not only in our education, but in the whole social system of our country ; and the interest of many other countries in this matter is not less than our own.

(iii) The investigations on examinations recorded in this book are the outcome of an International Conference on Examinations held in May, 1931, at Eastbourne, under the auspices of the Carnegie Corporation, the Carnegie Foundation, and the International Institute of Teachers College, Columbia University. The countries represented at the Conference were (in alphabetical order) England, France, Germany, Scotland, Switzerland, and the United States.[2] As a result of that Conference, committees were set up in all the European countries above-named. Each of these committees received a grant for

[1] In a Conspectus prepared for the Committee there appear over 160 names of such bodies, exclusive of Universities and Local Education Authorities.

[2] The Report of the Eastbourne Conference on Examinations, edited by Professor Paul Monroe, Director of the International Institute, was published by the Bureau of Publications, Teachers College, Columbia University, New York City, in 1931.
The representatives from the United States at the Conference were as follows :—
Dr. C. H. Judd, Dean of the School of Education, University of Chicago.
Dr. Frederick P. Keppel, President of the Carnegie Corporation, New York City.
Dr. Paul Monroe, Director of the International Institute, Teachers College, Columbia University.
Dr. Henry Suzzallo, President of the Carnegie Foundation, New York City.
Dr. Edward L. Thorndike, Professor of Education, Teachers College, Columbia University.

three years from the Carnegie Corporation through the International Institute, and each of them reported independently to a second International Conference held in June, 1935, at Folkestone, under the same auspices as the Conference held at Eastbourne. The Committees have done their work on independent lines and have reported separately.

(iv) The English Committee consisted of the following: Sir Michael Sadler, K.C.S.I. (Chairman), Dr. P. B. Ballard, Dr. C. Delisle Burns, Professor Cyril Burt, Sir Philip Hartog, K.B.E. (Director), Professor Sir Percy Nunn, Professor C. Spearman, F.R.S., and Professor Graham Wallas. The Committee suffered a great loss in 1932 by the death of Professor Graham Wallas, who was replaced by Professor Godfrey Thomson, a member of the Scottish Committee. Professor H. R. Hamley and Professor F. Clarke joined the English Committee later. Professor C. W. Valentine was elected a member in July, 1935, and resigned at the end of December in the same year.[1] The

[1] The membership of the other Committees is shown below:—

FRANCE—

M. A. Desclos, Directeur-adjoint de l'Office National des Universités et Écoles Françaises (*President*).

M. Barrier, Adjoint au Directeur de l'Enseignement Primaire.

M. Bouglé, Directeur de l'École Normale Supérieure.

M. Gastinel, Inspecteur Général de l'Instruction Publique.

M. Laugier, Maître de Conférences à la Faculté des Sciences de Paris.

M. Luc, Directeur-adjoint de l'Enseignement Technique.

The original Committee included:—

M. Charles Maurain, Doyen de la Faculté des Sciences de l'Université de Paris (who resigned on account of the pressure of other duties).

M. Cope, Président du Syndicat National des Professeurs des Lycées de Garçons et de l'Enseignement Secondaire Féminin (since deceased).

GERMANY—

Professor Erich Hylla, Ministerialrat im Ministerium für Kunst, Wissenschaft, und Volksbildung in Preussen; Professor an der Pädagogischen Akademie, Halle.

Dr. Robert Ulich, Ministerialrat im Ministerium für Volksbildung in Sachsen.

The original Committee included also:—

Professor Dr. Carl Becker, Minister a.D. für Kunst, Wissenschaft, und Volksbildung in Preussen; Professor an der Universität, Berlin (since deceased).

Dr. Otto Bobertag, University of Berlin (since deceased).

SCOTLAND—

William Boyd, M.A., B.Sc., D.Phil., Lecturer in Education, Glasgow University.

Shepherd Dawson, M.A., D.Sc., Lecturer in Psychology, Jordanhill Training College, Glasgow (since deceased).

Professor James Drever, M.A., D.Phil., Professor of Psychology, Edinburgh University.

Thomas Henderson, B.Sc., F.E.I.S., Hon. Secretary of the Scottish Council for Research in Education.

W. A. F. Hepburn, M.C., M.A., B.Ed., Director of Education to the Ayrshire Education Committee.

Professor W. W. McClelland, M.A., B.Sc., B.Ed., Professor of Education, St. Andrews University.

J. Mackie, M.A., D.Sc., F.R.S.E., Head Master, Leith Academy.

[*Continued next page*

address of the English Committee is 1, Plowden Buildings, Temple, London, E.C.4.

(v) Touching education and social life as they do on so many points, the problems of examinations are many and varied. The Committee have published an *English Bibliography of Examinations* (1900-32),[1] which shows how much has been written on the subject in this country during the first third of the century. They have also published a volume of *Essays on Examinations*, dealing with a number of aspects of the subject, and have prepared a *Conspectus of Examinations in Great Britain and Northern Ireland.*

A summary of the investigations described in the present volume was published under the auspices of the Committee in a pamphlet entitled *An Examination of Examinations*, by Sir Philip Hartog and Dr. E. C. Rhodes, early in December, 1935. A second impression was published shortly afterwards, and a second edition (third impression) in May, 1936.

(vi) The object of the investigations to be described in this volume may be explained very simply. Professor F. Y. Edgeworth, many years ago, found that the marks allotted independently by twenty-eight different examiners to a single piece of Latin prose varied from 45 to 100 per cent., and made a number of other investigations on variability in marking. In the United States, Messrs. Starch and Elliot, and, in France,

[1] *An English Bibliography of Examinations* (1900-1932), by Mary C. Champneys, with a Foreword by Sir Michael Sadler and Sir Philip Hartog (Macmillan & Co., Ltd.), 1934.

Robert R. Rusk, M.A., B.A., Ph.D., Lecturer in Education, Jordanhill Training College, Glasgow; Director to the Scottish Council for Research in Education.

J. C. Smith, C.B.E., M.A., D.Litt., formerly Senior Chief Inspector of Schools, Scottish Education Department.

Professor Godfrey H. Thomson, Ph.D., D.Sc., Professor of Education, Edinburgh University.

SWITZERLAND—

M. Pierre Bovet, Professeur à l'Université de Genève; Directeur de l'Institut Universitaire des Sciences de l'Éducation, Genève.

Dr. Brenner, Directeur du Lehrerseminar, Bâle.

M. Edouard Claparède, Professeur de Psychologie à l'Université de Genève; Directeur de l'Institut Jean-Jacques Rousseau.

M. Robert Dottrens, Directeur d'Écoles, Troinex, Genève (Dr. Soc.).

Dr. Charles Junod.

M. Albert Malche, Conseiller aux États; Professeur à l'Université de Genève.

M. Jean Piaget, Directeur du Bureau International d'Éducation, Genève; Professeur extraordinaire à l'Université de Genève; Co-directeur de l'Institut Jean-Jacques Rousseau.

Dr. W. Schohaus, Schweizerische Erziehungs Rundschau, Kreuzlingen, Thurgovie.

Dr. Ida Somazzi, Seminar, Berne.

Dr. Hans Stettbacher, Lehramtkurse, Universität, Zurich.

M. Teodoro Valentini, Professeur, Scuola Normale, Locarno, Tessin.

M. Laugier and Mlle. Weinberg, have found similar results,[1] but no systematic comparison has hitherto been published of the marks allotted by a number of different examiners and by different boards of examiners, all experienced and qualified for their task, to sets of scripts[2] actually written at public examinations. Both the English and the French Committees have attacked this subject. A brief summary of the French results is appended to *An Examination of Examinations* (pp. 78-81). The French Committee have published, besides their *Atlas de l'enseignement en France* (in quarto-raisin, pp. xiii, 183, 13 planches hors texte, 75 francs), a volume entitled *La correction des épreuves écrites dans les examens, enquête expérimentale sur le baccalauréat.* Both books are published by the Maison du Livre, 4 Rue Félibien, Paris. The results are similar in the two countries and equally disquieting.

(vii) In carrying out the investigations, the following general principles were observed :—

(1) The scripts investigated were all actual scripts which had been written by candidates in the course of an ordinary examination. It was only after long and delicate negotiations with the various bodies that the actual scripts could be secured.

(2) The scripts used were written as answers at the following examinations, which were chosen by the Committee as important and typical :—

(a) *School Certificate Examinations*, for which there are between 60,000 and 70,000 candidates every year. These are the School Leaving Examinations taking place at the age of about 16, the passing of which under certain conditions qualifies for entrance to a university and to a number of

[1] Professor F. Y. Edgeworth's statistical investigations on the results of examinations are contained in three memoirs: (i) The Statistics of Examinations, *Journal of the Royal Statistical Society*, vol. li (1888), pp. 599-635, (ii) The Element of Chance in Competitive Examinations, *ibid.*, vol. liii (1890), pp. 460-75 and 644-63, and (iii) On Problems in Probabilities (*Philosophical Magazine* for August, 1890). A summary of these memoirs, revised by Professor Edgeworth, is contained in Hartog's *Examinations*, etc., 1918. Further investigations were made by Starch and Elliot in the United States (see D. Starch, *Educational Psychology* (1920), p. 433, and the Bibliography; also Starch, *Educational Measurements* (1916), p. 3 *et seq.*), and by M. Laugier and Mlle. Weinberg in France (Le facteur subjectif dans les notes d'examen, *Année Psychologique*, xxvii (1927), pp. 236-44, and xxviii (1928), pp. 229-41; but their results, though striking and interesting, are on a relatively small scale.

[2] The technical term " script " is used to designate the book or books containing the answers of a single candidate to a paper of questions set at an examination. By extension, the term " script " is occasionally used (e.g., in Chapter II) to designate the books containing the answers of a single candidate to two or more papers on the same subject, for which the marks are added together for the purpose of the examination.

professions. A School Certificate is also required as a condition of engagement by many business men.

(b) *Special Place Examinations.* These are the examinations held for children between the ages of 10 and 12, on the results of which pupils in elementary schools at present gain admittance to central schools or secondary schools. The number of entries every year is estimated at from 400,000 to 500,000.

(c) *A College Scholarship Examination* at one of the older universities in *English Essay.*

(d) *A University Honours Examination in Mathematics.*

(e) *A University Honours Examination in History.*

(3) Every mark on the scripts made by the original examiners was completely removed before they were circulated or photographed.

(4) The examiners by whom the papers were marked (men and women) were in every case examiners with experience of the kind of examination investigated. In four of the investigations on School Certificate Examinations the examiners in the various subjects were chosen in each case from a large panel of a single examining body (other than the body which had supplied the scripts)[1], including both persons of the rank of Chief Examiner, and Assistant Examiners specially recommended on the ground of their experience and ability. The examiners for the College Entrance Scholarship Essay scripts and for the University Mathematical Honours scripts were in both cases examiners of the university for which the scripts were written. For the History Honours scripts it was impossible to secure a sufficient number of examiners from the same university, and the seventeen examiners concerned were chosen from nine different universities and included nine university professors.

(5) The time allowed for the correction of the scripts was, as a rule, the time desired by the examiners concerned. It may be fairly said that the scripts were corrected under less pressure in respect of time than ordinarily prevails at an examination, so that the marks may be regarded as expressing the deliberate opinion of the examiners concerned.

(6) Every examiner was furnished with a mark-sheet providing for separate entries for each answer, and in some cases for parts of an answer.

[1] In the investigation on School Certificate English conducted under the auspices of the Durham University School Examinations Board, of which the main results are reproduced in this volume, the examiners were not all chosen from the panel of the same examining body (see pp. 64-67 below).

(7) Every precaution was taken to ensure that no answer was overlooked by an examiner, and in any case of doubt the script was returned to the examiner for reconsideration.

(8) The examiners were all paid either in accordance with the usual scale adopted for the marking of scripts of the same kind, or, in certain cases, on a scale slightly higher. The Committee regard the payment of the examiners as an essential feature of the investigation. It might have been possible to secure the voluntary help of competent examiners, but marking carried out by voluntary helpers would have been carried out under conditions different from those of a real examination. In an investigation of this kind it is to be remembered that the actual task of marking examination scripts is for most examiners wearisome, and the psychological condition of a person who is unpaid for performing such work is likely to be different from the condition of a person who is adequately paid.

(9) The report of the investigations on each set of scripts was submitted to the original examining body before publication.

(10) The marks were all analysed by Dr. E. C. Rhodes, Reader in Statistics in the University of London, who has acted as statistician for the Committee and who co-operated in the investigations. The results were submitted to the Committee as each investigation terminated.

(11) The Committee are anxious that their investigations should not be interpreted as a criticism of any particular body. No mention has been made in these investigations of the marks allotted to the scripts by the original examining bodies.

(viii) The Committee believe that, in view of the precautions taken, the discrepancies between the marks of the different examiners afford an indication of the element of chance in examinations as they are at present conducted. The investigations show how a change in the selection of particular examiners, from a panel of persons who are all experienced and regarded as well qualified, would tend to affect the fate of individual candidates.

In Memorandum III below on School Certificate Examinations attention is called to the " machinery of examinations," called " standardisation " (see para. 658 below), and the great trouble taken by School Certificate Examination authorities to investigate " border-line " cases and to modify the results of an examination in a particular subject by other considerations, notably the general performance at the examination and the reports of

teachers, so that the chances of a candidate being "wrongly" rejected are materially diminished. Precautions are of course also taken by other authorities, e.g., those who conduct Free Place Examinations, in regard to border-line cases. But it must be pointed out that candidates may be placed in error either above or below the border-line. Thus a deserving candidate may by chance be deprived of a pass or credit; and, conversely, an examination certificate which is regarded by the public as a certificate of efficiency may by chance be given to candidates who by rights should be rejected. There are many examinations in which the final results in a particular subject may depend solely on the marks allotted by two examiners or even by a single examiner. Our results show how serious may be the element of chance in such cases.

(ix) We must guard ourselves here against the suggestion that the chances due to divergences of marking are the only ones in the examination system. There is also the element of chance due to the variability of condition in individual candidates, arising from illness or accident, which it is difficult to estimate statistically. It may be reduced, in a rough and ready way, when examining bodies take into account school-records in border-line cases. Then there is the element of chance due to variability in the difficulty of the papers set. This again some examining bodies dealing with large numbers of candidates attempt to reduce by correcting the marks assigned by the examiners to candidates, so as to make them conform (in accordance with a suggestion of Edgeworth) to a curve regarded from experience as being suitable for the particular examination. This last expedient helps to avoid violent fluctuations in the proportion of those who pass or fail, or are awarded marks of credit and distinction. We shall not here discuss this last kind of adjustment. We deal only with the question of the original differences of marks allotted by a number of examiners to the same scripts.

(x) Besides the investigations into written examinations, the Committee carried out one investigation of a particularly interesting nature into the consistency of the marking of two boards of examiners at an interview of the same kind as that held at Civil Service Examinations, with the object of testing such characteristics as "alertness, intelligence and intellectual outlook." (see p. 169 below). The investigation yielded remarkable results.

(xi) To place the investigations in their right perspective it is necessary to remind the reader that in examinations, as in

all other psychological tests, it is of importance to bear in mind both their " validity " and their " reliability," or " consistency." By " validity " is meant the degree of agreement of a measurement with the thing measured. By " reliability " is meant the degree of agreement between any two independent sets of measurements of the same set of things.

(xii) To quote Professor Spearman, " the inter-relations of reliability and validity are one-sided. Low reliability necessarily involves low validity, but the converse is not true. Wherever we find bad agreement between different measurements, then we can safely say that the examination is bad. But when the measurements agree we *cannot* forthwith say that the examination is good."[1]

(xiii) This book, as a whole, is devoted to the study of the " consistency " and not directly of the " validity " of examinations ; their " validity " is dealt with indirectly, since the low " consistency " which in certain cases we have found in marking necessarily involves low " validity." Looked at from another point of view, low consistency means that for individual candidates the element of chance in important examinations (which may determine a career) must be great. The question of consistency is therefore one of great practical importance.

(xiv) In Part I are given full details of the procedure and the numerical results of the eleven investigations recorded.

(xv) In Part II, and in the Memoranda by Professor Burt and Dr. Rhodes which follow it, the problem has been analysed psychologically and the results statistically. Dr. Rhodes (in paras. 365-545 and 608-633) and Professor Burt (in paras. 560 *et seq.*) envisage and analyse in somewhat different ways the fundamental causes which lead different examiners to assign differing, and sometimes widely differing, marks to the same piece of work.

(xvi) Again, both contributors aim by statistical methods at deducing from the data the " ideal mark " for each piece of work or script, and at classifying the examiners in order of merit, and so determining in respect of a given set of data which

[1] The word " consistency " will be used below as a substitute for the term " reliability," as used by Professor Spearman in his " Note on the Reliability and Validity of Measurements " in the *Essays on Examinations* published by the Committee (Macmillan, 1936), from which the passage quoted above is taken. It is easy to give an example of Professor Spearman's dictum. Let us suppose an examination of which the main purpose is to test the ability of a candidate to translate Latin prose into English. In such papers it is a common practice to include tests on accidence, of which the " consistency " is often high (see paras. 47-48, pp. 28-29 below) ; but, since an excellent memory for details of accidence may be associated with inability to translate Latin prose into English, the validity of the accidence test, in respect of the general purpose of the examination, is low.

is the "best examiner." It is recognised by the two contributors that different examiners may not only adopt different "standards" of marking, but that they may fail to adhere rigidly to those standards and hence introduce "random variations" into their marking. Dr. Rhodes gives a first approximation for obtaining his "ideal marks" in Part II, and gives a second approximation in his Memorandum II (pp. 315-324). It is worthy of remark that, in dealing with the present data, the results yielded by the different mathematical methods of Dr. Rhodes and Professor Burt are not significantly different (see para. 603 (3), p. 309). The Committee believe that, apart from the present numerical data, the methods described by the two contributors will be of more general application, and of solid interest to educational psychologists, in connexion with other problems.

(xvii) The term "best examiner," however, as the two writers fully recognise, cannot be considered solely on the basis of a statistical comparison of the marks awarded by a number of examiners dealing with the same scripts. It raises once more the question of the purpose of each particular examination. In a rationally conducted examination the purpose will be in the minds of the examiners at every stage :—

(a) it will control the construction of every question-paper and practical test ;[1]

(b) it will control the marking-scheme, i.e. the distribution of marks between the different portions of a question-paper (unless the paper is marked "by impression") ;

(c) it will control the allotment of marks within the marking-scheme.[2]

On the other hand, uncertainty of purpose or the presence of conflicting purposes will introduce uncertainty at every stage. (See Memorandum III, p. 325 below.)

[1] The decision of the Committee to conduct their investigations with the use of scripts that had been actually written in the examination-room excluded the possibility of making experiments with different question-papers ; since the control of the question-papers naturally rested with the examination authorities from whom the scripts were obtained. But, as will be seen below, the Committee have borne in mind the desirability of experimenting with the use of different question-papers at a later stage.

[2] The great divergencies between the marks of individual examiners assigned to identical answers in translation into French, within the limits of a marking-scheme, show how much scope there is for the exercise of individual judgment by examiners in this matter (see Table 28, p. 49 below). It is conceivable that an examiner whose results may be divergent from the average, or even erratic, as judged by Chief Examiners, may have the right to be regarded as a better examiner than his chief or his colleagues if he has a clearer realisation in his mind of the purpose of the examination and a finer perception of the quality of the candidates.

(xviii) The examinations of which the purpose is the most clearly defined are technical examinations of which the aim is to determine whether a candidate has or has not a definite utilisable skill. Certain statistical considerations which, as we shall see, play a great part in some other examinations, are reduced here to a minor role. When qualifying examinations are held on the work of an actuary, an accountant, a surgeon, an air-pilot, a motor-driver, or a shorthand-typist, experienced examiners are as ready to examine a single candidate as to examine fifty, and to say whether he is or is not suitable for his particular job. If there are, say, twenty candidates, the examiners may pass or plough the whole twenty. The examiners have an acute sense of responsibility to the general public. The exact percentage of marks required for a " pass " is here of comparatively little consequence. Where, as is generally the case, a test is a complex one made up of a number of tests, it is the duty of examiners to see that if, say, by an accumulation of marks for points of minor importance, that percentage has been allotted in the first instance to a candidate judged on the whole to be unfit for his job, it shall be reduced below the border-line for a pass.[1]

(xix) Below the border-line (except for the information of the candidate, who rarely receives it) the marks in such examinations are of no importance. And above the border-line they only become of importance if the examiners wish to classify or compare the performances of different candidates, all of whom are fit to do the required job; in other words, when the examination becomes competitive. Candidates may of course be placed in order " by impression." In complex examinations, covering different subjects, the candidates are more frequently placed in order by the addition of marks. How far this process of addition is theoretically justifiable, without a careful scrutiny made in the light of the general purpose of the examination, is a matter for consideration in each special case. (It will be seen later that certain large-scale examinations, in which the fate of a candidate depends upon a sum of marks and which are not generally regarded as competitive, may be so in reality.)

[1] The percentage for a pass may be placed so high that an adjustment of this kind would rarely be necessary. In the General Information and Conditions of Examination for Civil Air Navigation Licences (Second Class), A.M. pamphlet 44 (3rd edit., Jan., 1934), it is stated that:—

" In order to qualify, candidates will be required to obtain not less than 90 per cent. of the total marks in visual signalling, not less than 60 per cent. of the marks in any other subject, and not less than 70 per cent. of the total marks in all subjects, excluding visual signalling."

(xx) In sharp contrast with examinations which test technical efficiency are certain important "general" examinations such as the School Certificate Examinations referred to above, which may test certain "utilisable skills," but which, over a large field, must be regarded only as tests of progress towards the attainment of such skills, and in which the progress of the different candidates is compared.

(xxi) A simple example will illustrate what we mean. Let us suppose that, at some elementary examination, examiners are comparing the performance of two candidates in simple addition. Candidate A gets (say) 6 sums out of 20 right, and Candidate B gets 8 sums out of 20 right. In an examination which is a test of progress B should have more marks than A. But neither candidate can add; neither has a "utilisable skill" in addition. Whatever the scheme of marking, common-sense demands that no certificate given to either candidate should imply that he "can add." Here the decision is clear cut. The candidates have failed in addition. But when examiners are testing (say) History at the School Certificate stage, what is to be the criterion? What is a "minimum pass" to mean in terms of performance? What is it to mean in English or French, or Latin or Chemistry?

(xxii) Memorandum III on School Certificate Examinations appended to this book shows clearly how much confusion may occur with the present system, both at the border-line of pass and at the border-line of credit. If the test at the border-line could be regarded as a test of a utilisable skill, decisions would be comparatively simple and the allotment of marks, as shown above, relatively unimportant. But a pass interpreted in terms of progress is another matter. If we are comparing candidates in terms of progress only, it may be justifiable for examiners to say, as in School Certificate History (Memorandum III, paras. 647-648) that they will "pass" the 75 per cent. of the candidates who get the highest marks (however bad some of the candidates may be from the point of view of performance) and plough the rest. The examination is a competitive test of progress and not of attainment. There is therefore a serious, though not an obvious, competition for a "pass," and one in which consistency of marking is of great importance to the individual candidate.

(xxiii) The employment of boards of examiners instead of individual examiners, though it diminishes, does not remove the element of chance in examinations, and boards, as well as individuals, using the same examination-papers, may disagree in their verdicts (see Chapters III and IV).

(xxiv) The question may be asked: Should examinations be abolished? If not, what remedies can be suggested?

The Committee are clearly opposed to the root and branch policy, on the ground that examinations as a test of efficiency are necessary. They are of opinion that a more extended use might be made of examinations which yield identical results when applied by different examiners (e.g., "New Type" or other "Objective" examinations), but that the traditional "essay" examination should be preserved, because it tests, though at present with considerable uncertainty, skills which cannot be tested by "new-type" tests, e.g., the power to present a complex series of facts or arguments. But they hold that it is as impracticable to recommend an *a priori* cure for the defects of the present examination system as it would be to recommend an *a priori* cure for a disease. It is only by careful and systematic experiment that methods of examination can be devised not liable to the distressing uncertainties of the present system.[1] No doubt investigations like those undertaken by the Committee, and administrative experiments in allowing teachers, in conjunction with Government or University inspectors, to "brand their own herrings" would involve expenditure, but such expenditure and experiments would be justified in the public interest.

(xxv) The Committee desire to acknowledge their deep obligation to the various examination authorities by whom they have been furnished with the scripts which formed the material for their investigations, or by whom they have been assisted in other ways, and to the examiners who marked the scripts or took part in the *viva voce* examination. Without the cordial assistance both of examination authorities and of examiners, it would have been impossible for the Committee to carry out their investigations on the lines which they had planned.[2] They are especially indebted to Dr. Rhodes for the skilled ability and indefatigable zeal which he has devoted to the work of the Committee. They also desire to acknowledge

[1] With the help of a fresh grant for 1936 the Committee are making experiments with a view to improving both the validity and the consistency of tests in English composition. If such experiments prove successful, they may conceivably lead to a general improvement of the essay type of examination.

[2] They also wish to acknowledge their indebtedness to Mr. John Bell, the High Master of St. Paul's School, who gave them the necessary facilities for a comparison of the marks earned by pupils when they were examined in History in the ordinary way with those earned when they were allowed the use of textbooks for reference at the examination. The results were of too fragmentary a character to justify publication at this stage.

the efficient and devoted assistance given to them in the secretarial and checking work by Miss Gladys Roberts and Mr. E. C. Rubidge.

In conclusion, the Committee wish to express their warm appreciation of the generosity and initiative of the Carnegie Corporation, the Carnegie Foundation, and the International Institute of Teachers College, Columbia University, to which this Committee and the parallel Committees in other countries owe their existence ; and in this connexion it should be mentioned that the moving spirit in the organisation of the two Conferences has been that veteran in international education, Dr. Paul Monroe, whose devoted service to the work of the international committees it is impossible to exaggerate.

For the Committee,

M. E. SADLER, *Chairman.*

P. J. HARTOG, *Director.*

April, 1936.

THE MARKS OF EXAMINERS

PART I

By
P. J. Hartog and E. C. Rhodes

CHAPTER I

MARKING OF SCHOOL CERTIFICATE HISTORY SCRIPTS (TWO INVESTIGATIONS)

1. *Character of the Examination Paper.*—The paper contained over twenty questions on Modern History, of which six were to be answered at the choice of the candidate. The maximum mark fixed for each question was 16. The time allowed was three hours.

2. *Special Objects of the two Investigations and Method of Selection of Scripts.*—The objects of the two investigations were as follows :—

(i) to examine the discrepancies of different examiners in marking scripts which, on a first marking, appeared to be of equal value. For this purpose fifteen scripts were selected which had been awarded exactly the same "middling" mark by the School Certificate authority concerned. (As in other cases, every trace of the origin of the script and original marking was removed from each script.)

(ii) to compare the marks allotted by a number of individual examiners to the same series of scripts on two occasions separated by a sufficient interval of time to ensure that the examiners would not on the second occasion recollect the marks which they had given on the first.

3. First Investigation. *Procedure.*—In order to secure examiners accustomed to team-work, fifteen were selected from the panel of a single School Certificate authority. This authority was not the one by which the scripts were furnished. The examiners were informed that the object of the investigation was to form an estimate of the differences in marking between different

examiners, but they were not informed that the scripts had been originally marked as being of the same value ; nor was any other indication given of the original mark assigned to them. They were all asked to assign awards of Failure, Pass or Credit to each script and to state what limiting mark they fixed for these awards.

A Board of Examiners was expressly dispensed with in this investigation. Each examiner was treated as if he were a single head-examiner ; and it was on this account that he was left free to decide his own limiting numerical marks for the awards, on the basis of his own experience. It will be seen from paragraph 5 below that the differences of these numerical limits are not negligible. They varied from 32 to 42 for a Pass, from 45 to 50 for a Credit. The difference between the minimum mark for a Pass and the minimum for Credit was fixed by the individual examiners as follows :—

6 marks	Examiner L
10 marks	Examiners B, C, E, F, G, J, K, M, N, P, Q
13 marks	Examiner D
16 marks	Examiners A and H

It is to be noted that eleven examiners made the range 10 marks, which we believe is adopted in many cases by School Certificate authorities.[1] The fixing of a range between Pass and Credit as low as 6 or as high as 16 marks may perhaps be regarded as exceptional.

The examiners were informed that 16 marks was the maximum for each question.

4. SECOND INVESTIGATION. *Procedure.*—After an interval which varied with the different examiners, but was not less than 12 or more than 19 months in any case,[2] the same examiners were requested to re-examine the scripts with a view to a comparison of their second marking with their first marking ; and they were informed of the object of the investigation. Fourteen of the fifteen examiners who acted in the first investigation took part in the second investigation, the exception being Examiner A. These fourteen assured us that they had kept no record of their

[1] The Report of the Investigators on the School Certificate Examination, published in 1932 (H.M. Stationery Office), pp. 31-32, states that : " There is a fairly definite relation between the credit and the pass marks in a subject (though as explained these are not necessarily fixed marks) ; thus in one large examination the pass mark is four-fifths and in another seven-ninths of the credit mark. . . ." If the Credit mark were approximately 50, this would correspond to a Pass mark of approximately 40.

[2] The scripts had to be circulated in an order which was not identical with the original order and was determined by the convenience of the examiners.

previous work. The scripts were renumbered in an entirely different and unrelated order and again submitted to the examiners with instructions identical with those given in the first instance.[1]

5. *Principal Results of the First Investigation.*—Table 1 below shows the marks assigned by the fifteen examiners A to Q to the fifteen scripts and the limits fixed by each examiner for Pass and Credit.

TABLE 1

TOTAL MARKS OF THE FIRST INVESTIGATION (UNADJUSTED)

No. of Candidate	Examiner															Range, i.e. difference between highest and lowest mark
	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	
1	38	31	43	33	30	34	44	46	31	48	43	42	31	41	50	20
2	38	26	37	32	41	37	29	43	31	32	34	41	32	34	42	17
3	33	29	55	35	40	43	41	49	43	51	40	46	36	45	46	26
4	39	38	52	36	63	61	48	52	52	53	40	52	50	39	43	27
5	42	32	41	33	42	47	40	45	47	48	42	39	32	38	48	16
6	27	35	37	48	53	53	62	51	41	46	38	49	48	50	47	35
7	48	46	53	36	57	48	60	58	58	53	46	45	53	50	51	24
8	38	21	45	27	37	41	39	36	36	39	35	41	25	38	45	24
9	30	26	37	30	27	29	38	44	43	38	29	40	24	24	37	20
10	33	39	39	35	39	40	50	55	57	52	42	47	45	38	40	24
11	34	28	50	37	48	48	41	49	55	49	36	46	38	41	52	27
12	52	38	68	35	60	48	49	61	70	69	43	53	44	57	62	35
13	24	24	35	30	34	31	32	40	38	41	37	43	38	40	45	21
14	32	39	30	33	40	44	58	53	47	44	43	41	34	27	41	31
15	53	42	55	47	45	70	60	65	53	58	51	53	52	49	53	28
Pass mark	34	38	40	32	40	38	38	34	40	38	42	40	40	38	40	
Credit mark	50	48	50	45	50	48	48	50	50	48	48	50	50	48	50	

6. Before entering into detail, we may point out a salient feature in this Table. Whereas the scripts had originally been all allotted the same moderate mark, they were allotted by these fifteen examiners 42 different marks varying from 21 to 70. It is to be noted that the maximum is 96 and not 100.

7. It might fairly be suggested that some of these surprising differences were due to the different numerical limits fixed by the different examiners for Failure, Pass, and Credit. We have therefore adjusted the marks of each examiner so as to assign in all cases the mark 40 to the scripts to which he assigned a bare Pass, and 50 to the scripts to which he assigned a bare Credit. Neither the order of the candidates, nor the award of

[1] It seems beyond the range of probability that any examiner should have had any recollection of the mark previously assigned by him or her to a particular script. As will be seen, the results show no evidence of any such recollection.

Failure, Pass and Credit to any candidate is altered by our method of adjustment, which is explained in the footnote below.[1]

The adjusted marks are set out in Table 2 below :—

TABLE 2

ADJUSTED MARKS OF THE FIRST INVESTIGATION

No. of Candidate	Examiner															Range, i.e., difference between highest and lowest mark
	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	
1	43	33	43	41	30	36	46	48	31	50	42	42	31	43	50	20
2	43	27	37	40	41	39	31	46	31	34	32	41	32	36	42	19
3	39	31	55	42	40	45	43	49	43	53	38	46	36	47	46	24
4	43	40	52	43	63	62	50	52	52	55	38	52	50	41	43	25
5	45	34	41	41	42	49	42	47	47	50	40	39	32	40	48	18
6	32	37	37	53	53	55	63	51	41	48	36	49	48	52	47	31
7	49	48	53	43	57	50	62	58	58	55	47	45	53	52	51	19
8	43	22	45	34	37	43	41	41	36	41	33	41	25	40	45	23
9	35	27	37	38	27	31	40	46	43	40	28	40	24	25	37	22
10	39	41	39	42	39	42	52	55	57	54	40	47	45	40	40	18
11	40	29	50	44	48	50	43	49	55	51	34	46	38	43	52	26
12	52	40	68	42	60	50	51	61	70	70	42	53	44	59	62	30
13	28	25	35	38	34	33	34	44	38	43	35	43	38	42	45	20
14	38	41	30	41	40	46	60	53	47	46	42	41	34	28	41	32
15	53	44	55	52	45	71	62	65	53	60	53	53	52	51	53	27

Pass Mark 40
Credit Mark 50
} for all examiners.

As compared with the unadjusted marks of Table 1, we find that the lowest mark is increased from 21 to 22, and the highest from 70 to 71, the extreme range remaining the same. The remarkable differences in the estimates of the examiners persist.

8. *Order of the Candidates.*—The differences between the numerical estimates of the examiners are demonstrated in another way in Table 3 below, which shows the " order of merit " in which the candidates are placed by them.[2]

[1] The process may be illustrated by taking the marks of Examiner D, who fixed 32 for a Pass and 45 for a Credit.

The original mark of 32 was converted into 40 ; all marks below 32 were multiplied by $\frac{5}{4}$ so that an original mark 27 became 34 after adjustment.

Marks from 32 to 45 were adjusted by adding to 40 the difference between the original mark and 32 multiplied by $\frac{10}{13}$. Thus 45 became $40 + (13 \times \frac{10}{13}) = 50$; and 36 became $40 + (4 \times \frac{10}{13}) = 43$.

A mark above 45 was subtracted from 96, the difference multiplied by $\frac{10}{51}$, and this difference again subtracted from 96. Thus the original mark 47 yielded an adjusted mark 52 [$96 - 47 = 49$; $49 \times \frac{10}{51} = 44$; $96 - 44 = 52$].

[2] We have used the usual term " order of merit " here and later in the book, though the term " order of proficiency " seems to us more suitable.

TABLE 3

Order of Merit

Candidate \ Examiner	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q
1	7	9	8	10	14	13	8	10	14½	8½	4	10	13	6½	5
2	7	12½	12	12	8	12	15	13	14½	15	14	12	11½	13	12
3	10½	10	2½	7	9½	9	9½	8½	9½	6	8½	6½	9	5	8
4	5	5½	5	4½	1	2	7	6	6	3½	8½	3	3	9	11
5	4	8	9	10	7	7	11	11	7½	8½	6½	15	11½	11	6
6	14	7	12	1	4	3	1	7	11	10	10	4	4	2½	7
7	3	1	4	4½	3	5	2½	3	2	3½	2	8	1	2½	4
8	7	15	7	15	12	10	12	15	13	13	13	12	14	11	9½
9	13	12½	12	13½	15	15	13	12	9½	14	15	14	15	15	15
10	10½	3½	10	7	11	11	5	4	3	5	6½	5	5	11	14
11	9	11	6	3	5	5	9½	8½	4	7	12	6½	7½	6½	3
12	2	5½	1	7	2	5	6	2	1	1	4	1½	6	1	1
13	15	14	14	13½	13	14	14	14	12	12	11	9	7½	8	9½
14	12	3½	15	10	9½	8	4	5	7½	11	4	12	10	14	13
15	1	2	2½	2	6	1	2½	1	5	2	1	1½	2	4	2

There are striking differences exhibited in this Table, which can easily be noted by reading horizontally along each row.

Candidate														
1	is	mainly	placed	in	the	lower	half,	&	is	as	low	as	14½	& as high as 4
2	,,	,,	,,	,,	,,	,,	,,	,,	,,				15	,, ,, 7
3	,,	,,	,,	,,	,,	,,	,,	,,	,,				10½	,, ,, 2½
4	,,	,,	,,	,,	upper	,,	,,	,,	,,				11	,, ,, 1
5	,,	,,	,,	,,	lower	,,	,,	,,	,,				15	,, ,, 4
6	,,	equally in the lower half and the upper half					,,	,,	,,				14	,, ,, 1
7	,,	mainly	placed	in	the	upper	half	,,	,,	,,			8	,, ,, 1
8	,,	,,	,,	,,	lower	,,	,,	,,	,,				15	,, ,, 7
9	,,	always	,,	,,	,,	,,	,,	,,	,,				15	,, ,, 9½
10	,,	mainly	,,	,,	upper	,,	,,	,,	,,				14	,, ,, 3
11	,,	,,	,,	,,	,,	,,	,,	,,	,,				12	,, ,, 3
12	,,	always	,,	,,	,,	,,	,,	,,	,,				7	,, ,, 1
13	,,	,,	,,	,,	lower	,,	,,	,,	,,				15	,, ,, 7½
14	,,	mainly	,,	,,	,,	,,	,,	,,	,,				15	,, ,, 3½
15	,,	always	,,	,,	upper	,,	,,	,,	,,				6	,, ,, 1

Perhaps the most striking case is that of Candidate No. 6, who is placed 1st by two examiners, equal 2nd by one examiner, 3rd by one examiner, 4th by three examiners, 7th by three examiners, 10th by two examiners, 11th by one examiner, 12th by one examiner, and 14th by one examiner.

On the other hand, there is considerable agreement as to the good quality of Candidates No. 12 and No. 15, and as to the poor quality of Candidates No. 9 and No. 13.

9. *Classification of the Candidates.*—The following Table shows the number of Failures, Passes, and Credits allotted by the different examiners.

TABLE 4

CLASSIFICATION

Examiners	Failures	Passes	Credits
A	6	7	2
B	9	6	0
C	6	3	6
D	3	10	2
E	5	6	4
F	4	5	6
G	2	6	7
H	0	8	7
J	4	5	6
K	1	5	9
L	8	6	1
M	1	11	3
N	9	3	3
P	3	8	4
Q	1	9	5

Examiner H " ploughs " not a single candidate and awards eight Passes and seven Credits, while Examiner N " ploughs " nine candidates and awards three Passes and three Credits. Examiner L " ploughs " eight candidates and awards six Passes and only one Credit. Examiner K " ploughs " only one candidate and awards five Passes and nine Credits.

10. Table 5 shows the various awards to individual candidates :—

TABLE 5

Candidate	Examiner															Examiners' awards[1]		
	A	B	C	D	E	F	G	H	J	K	L	M	N	P	Q	F	P	C
1	P	F	P	P	F	F	P	P	F	C	P	P	F	P	C	5	8	2
2	P	F	F	P	P	F	F	P	F	F	F	P	F	F	P	9	6	0
3	F	F	C	P	P	P	P	P	P	C	F	P	F	P	P	4	9	2
4	P	P	C	P	C	C	C	C	C	C	F	C	C	P	P	1	5	9
5	P	F	P	P	P	P	P	P	P	C	P	F	F	P	P	3	11	1
6	F	F	F	C	C	C	C	C	P	P	F	P	P	C	P	4	5	6
7	P	P	C	P	C	C	C	C	C	C	P	P	C	C	C	0	5	10
8	P	F	P	F	F	P	P	P	F	P	F	P	F	P	P	6	9	0
9	F	F	F	F	F	F	P	P	P	P	F	P	F	F	F	10	5	0
10	F	P	F	P	F	P	C	C	C	C	P	P	P	P	P	3	8	4
11	P	F	C	P	P	C	P	P	C	C	F	P	F	P	C	3	7	5
12	C	P	C	P	C	C	C	C	C	C	P	C	P	C	C	0	4	11
13	F	F	F	F	F	F	F	P	F	P	F	P	F	P	P	10	5	0
14	F	P	F	P	P	P	C	C	P	P	P	P	F	F	P	4	9	2
15	C	P	C	C	P	C	C	C	C	C	C	C	C	C	C	0	2	13

[1] The letters F, P, C are used as abbreviations for Failure, Pass, and Credit respectively.

There is no complete agreement between the whole body of examiners in regard to the classification of any one candidate. They approach agreement most nearly in the case of Candidate No. 15, to whom thirteen examiners award Credit and two a Pass. Candidate No. 6 is " ploughed " by four examiners, awarded a Pass by five, and a Credit by six examiners. There are eight candidates altogether out of the fifteen who might in a real examination have received a failure-mark, a pass-mark, or a credit-mark, depending on the particular examiner who happened to mark his script.

11. We can pursue the origin of these differences by investigating the number of marks allotted by the different examiners to the answers to individual questions. They bring out the same kinds of difference.

The number of options was a large one, and the maximum number of candidates who answered any one question was ten.

Tables 6, 7 and 8 show the marks awarded for the answers to Qns. 13, 19 and 23.

TABLE 6

QUESTION 13

Candidate	2	14
Examiner	Marks Allotted	
A	10	0
B	7	2
C	9	3
D	8	3
E	13	3
F	8	4
G	5	7
H	7	7
J	7	3
K	5	3
L	7	4
M	10	4
N	9	2
P	9	4
Q	14	2
Lowest Mark	5	0
Highest Mark	14	7

TABLE 7

QUESTION 19

Candidate	1	3	4	5	8	9	10	11	12	13
Examiner	Marks Allotted									
A	8	6	6	8	7	8	4	4	12	4
B	7	5	4	5	4	3	2	2	9	7
C	9	9	7	7	8	5	4	6	14	9
D	8	5	3	7	4	4	4	4	8	8
E	7	4	12	6	7	5	3	5	14	11
F	8	7	10	10	6	6	3	5	13	8
G	9	9	8	7	7	8	3	4	10	7
H	10	8	7	8	6	8	5	4	13	12
J	6	7	9	8	8	6	6	4	15	12
K	12	10	6	9	7	4	3	6	13	10
L	10	9	4	7	7	4	4	5	9	10
M	9	7	6	7	7	6	4	5	10	10
N	8	5	4	4	4	2	2	5	8	8
P	10	10	4	10	7	4	3	5	13	13
Q	10	14	5	8	6	7	8	6	14	8
Lowest Mark	6	4	3	4	4	2	2	2	8	4
Highest Mark	12	14	12	10	8	8	8	6	15	13

TABLE 8

QUESTION 23

Examiner	Candidate 3	4	10	11	12	13	15
	Marks Allotted						
A	4	10	5	3	8	4	12
B	4	9	4	3	6	2	7
C	7	11	6	8	15	2	9
D	8	13	6	8	4	2	9
E	5	14	11	8	11	3	7
F	5	13	7	5	7	2	10
G	6	12	8	7	9	1	10
H	7	11	8	6	15	4	11
J	6	12	10	9	15	3	8
K	7	11	8	5	11	2	10
L	6	11	6	7	8	5	9
M	8	12	6	6	10	4	11
N	6	13	7	4	4	2	8
P	5	10	6	4	10	3	11
Q	5	12	7	7	8	6	9
Lowest Mark	4	9	4	3	4	1	7
Highest Mark	8	14	11	9	15	6	12

12. If, taking 40 as the Pass-mark and 50 as the Credit-mark out of a total of 96, we may regard 5 or 6 marks, for a question with a maximum of 16, as a Pass-mark, 8 marks as a Credit-mark, and 12 as " very good marks," a glance at the foregoing Tables will indicate the differences of view of different examiners in regard to particular answers.

Let us take the answer of Candidate No. 12 to Qn. 23. Three examiners (C, H and J) regarded this answer as of such excellence that they awarded it nearly full marks (15 out of 16); two (D and N) thought it so poor that they awarded it only 4 marks—25% of the maximum.

Take again the answer of Candidate No. 14 to Qn. 13; Examiner A regarded it as worthless; Examiners B, N and Q only gave it 2 marks, but Examiners G and H gave it 7, or nearly half-marks, a valuable contribution towards the number required for a Pass or a Credit.

13. We now turn to the candidates in regard to whom there is more agreement among the examiners. Let us compare the answers of Candidates No. 11 and No. 12 to Qn. 19. The marks allotted to these two candidates are absolutely different; they do not overlap. The marks of Candidate No. 11 range from 2 to 6, those of No. 12 from 8 to 15. All the examiners regard the first answer as poor or mediocre, and the second as (say) from " fair to good." The difficulty arises when the examiner tries to translate such epithets into marks. The question may

certainly be asked how far this is possible in a subject like History at the stage of the School Certificate Examination.

We know from experience that many experienced examiners in History object on principle to allotting numerical marks, and prefer the " literal " marks, α, β, γ, etc., of the " Oxford system." We shall discuss later the results of an investigation on History scripts (though of a far higher standard) in which literal marks were used (see Chapter X).

14. *Results of Second Investigation and Comparison with the Results of the First Investigation.*—We now turn to the second marking given by fourteen out of fifteen of the same examiners a year later (see para. 2 above), set out in Table 9 below :—

TABLE 9

SECOND MARKING. TOTAL MARKS AWARDED (UNADJUSTED)

	Examiner													
Candidate	B	C	D	E	F	G	H	J	K	L	M	N	P	Q
1	29	43	32	55	40	43	53	61	40	43	44	36	41	40
2	27	34	23	43	34	41	35	35	29	29	40	34	32	36
3	31	48	29	36	39	49	46	45	48	39	35	34	49	44
4	45	49	34	57	51	45	38	39	46	33	43	49	40	40
5	27	45	24	49	40	52	48	45	42	38	45	40	39	50
6	48	27	31	49	51	66	56	50	44	43	52	40	53	45
7	43	54	37	56	43	71	55	53	44	42	50	48	57	50
8	34	33	25	47	36	31	35	31	34	30	37	30	39	48
9	30	30	21	20	22	34	43	35	32	25	35	28	24	27
10	41	46	32	40	37	44	42	37	49	39	39	42	43	43
11	29	43	27	38	40	45	48	39	44	30	43	36	40	42
12	40	61	24	63	41	46	57	62	63	28	45	53	53	54
13	29	32	16	32	25	42	41	33	38	29	40	40	37	38
14	40	35	28	29	33	61	46	48	42	34	40	43	26	28
15	49	50	51	56	54	52	48	50	51	43	48	41	48	37
For Second Investigation :[1]														
Pass Mark	38	34	33	40	40	38	36	40	38	38	38	35	38	40
Credit Mark	48	45	50	50	50	46	48	50	48	48	48	49	48	50
For First Investigation :[1]														
Pass Mark	38	40	32	40	38	38	34	40	38	42	40	40	38	40
Credit Mark	48	50	45	50	48	48	50	50	48	48	50	50	48	50

15. The general character of the results, considered as a whole, and without reference to differences in regard to individual candidates, is the same as on the previous occasion (see paras. 5 and 6 above). The total number of different marks allotted to the fifteen candidates, who, it is to be remembered, were adjudged as of equal merit by the School Certificate authority,

[1] These limits were fixed by the examiner himself in each case (see para. 3 above).

was 44 on the second occasion, as against 42 on the first, and the extreme range of marks allotted was from 16 to 71 on the second occasion as against a range of from 21 to 70 on the first occasion. It is to be remembered that the number of examiners was one less on the second than on the first occasion—fourteen instead of fifteen. It will be seen that the differences of opinion between these fifteen examiners as to the merits of the various candidates would make any criticism of the original authority out of place. What has been demonstrated is not the failure of an individual or an examining authority but a failure in method, in all probability characteristic of any examination of this kind.

16. In order to provide for the differences of standard for Credit and Pass adopted by the different examiners, the marks awarded on the second occasion were adjusted in exactly the same way as that described in para. 7 above. It will be remembered that this adjustment does not alter either the order of the candidates, or the number of Failures, Passes, and Credits, assigned by the different examiners, but provides for a more accurate comparison of the numerical marks awarded, taking into account the differences of numerical marks allotted for the limits of Failure, Pass, and Credit.

17. Table 10 below shows the marks awarded by the fourteen examiners B to Q on the two occasions, as adjusted, and the differences between each pair of marks :—

TABLE 10

COMPARISON OF ADJUSTED MARKS OF FIRST AND SECOND INVESTIGATIONS

Candidate	Examiner B			Examiner C			Examiner D			Examiner E		
	1st Marking	2nd Marking	Difference	1st Marking	2nd Marking	Difference	1st Marking	2nd Marking	Difference	1st Marking	2nd Marking	Difference
1	33	31	−2	43	48	+5	41	39	−2	30	55	+25
2	27	28	+1	37	40	+3	40	28	−12	41	43	+2
3	31	33	+2	55	53	−2	42	35	−7	40	36	−4
4	40	47	+7	52	54	+2	43	41	−2	63	57	−6
5	34	28	−6	41	50	+9	41	29	−12	42	49	+7
6	37	50	+13	37	32	−5	53	38	−15	53	49	−4
7	48	45	−3	53	58	+5	43	42	−1	57	56	−1
8	22	36	+14	45	39	−6	34	30	−4	37	47	+10
9	27	32	+5	37	35	−2	38	25	−13	27	20	−7
10	41	43	+2	39	51	+12	42	39	−3	39	40	+1
11	29	31	+2	50	48	−2	44	33	−11	48	38	−10
12	40	42	+2	68	64	−4	42	29	−13	60	63	+3
13	25	31	+6	35	38	+3	38	19	−19	34	32	−2
14	41	42	+1	30	41	+11	41	34	−7	40	29	−11
15	44	51	+7	55	55	0	52	51	−1	45	56	+11
Average	34·6	38·0	+3·4	45·1	47·1	+1·9	42·3	34·1	−8·1	43·7	44·7	+0·9
Average difference, neglecting signs			4·9			4·7			8·1			6·9

TABLE 10—*continued*

Candidate	Examiner F 1st Marking	2nd Marking	Difference	Examiner G 1st Marking	2nd Marking	Difference	Examiner H 1st Marking	2nd Marking	Difference	Examiner J 1st Marking	2nd Marking	Difference
1	36	40	+4	46	46	0	48	55	+7	31	61	+30
2	39	34	−5	31	44	+13	46	39	−7	31	35	+4
3	45	39	−6	43	53	+10	49	48	−1	43	45	+2
4	62	51	−11	50	49	−1	52	42	−10	52	39	−13
5	49	40	−9	42	56	+14	47	50	+3	47	45	−2
6	55	51	−4	63	68	+5	51	58	+7	41	50	+9
7	50	43	−7	62	73	+11	58	57	−1	58	53	−5
8	43	36	−7	41	33	−8	41	39	−2	36	31	−5
9	31	22	−9	40	36	−4	46	46	0	43	35	−8
10	42	37	−5	52	48	−4	55	45	−10	57	37	−20
11	50	40	−10	43	49	+6	49	50	+1	55	39	−16
12	50	41	−9	51	50	−1	61	59	−2	70	62	−8
13	33	25	−8	34	45	+11	44	44	0	38	33	−5
14	46	33	−13	60	64	+4	53	48	−5	47	48	+1
15	71	54	−17	62	56	−6	65	50	−15	53	50	−3
Average	46·8	39·1	−7·7	48·0	51·3	+3·3	51·0	48·7	−2·3	46·8	44·2	−2·6
Average difference, neglecting signs			8·3			6·5			4·7			8·7

Candidate	Examiner K 1st Marking	2nd Marking	Difference	Examiner L 1st Marking	2nd Marking	Difference	Examiner M 1st Marking	2nd Marking	Difference	Examiner N 1st Marking	2nd Marking	Difference
1	50	42	−8	42	45	+3	42	46	+4	31	41	+10
2	34	31	−3	32	31	−1	41	42	+1	32	39	+7
3	53	50	−3	38	41	+3	46	37	−9	36	39	+3
4	55	48	−7	38	35	−3	52	45	−7	50	50	0
5	50	44	−6	40	40	0	39	47	+8	32	44	+12
6	48	46	−2	36	45	+9	49	54	+5	48	44	−4
7	55	46	−9	47	44	−3	45	52	+7	53	49	−4
8	41	36	−5	33	32	−1	41	39	−2	25	34	+9
9	40	34	−6	28	26	−2	40	37	−3	24	32	+8
10	54	51	−3	40	41	+1	47	41	−6	45	45	0
11	51	46	−5	34	32	−2	46	45	−1	38	41	+3
12	70	64	−6	42	29	−13	53	47	−6	44	54	+10
13	43	40	−3	35	31	−4	43	42	−1	38	44	+6
14	46	44	−2	42	36	−6	41	42	+1	34	46	+12
15	60	53	−7	53	45	−8	53	50	−3	52	44	−8
Average	50·0	45·0	−5·0	38·7	36·9	−1·8	45·2	44·4	−0·8	38·8	43·1	+4·3
Average difference, neglecting signs			5·0			3·9			4·3			6·4

TABLE 10—*continued*

Candidate	Examiner P 1st Marking	2nd Marking	Difference	Examiner Q 1st Marking	2nd Marking	Difference
1	43	43	0	50	40	−10
2	36	34	−2	42	36	−6
3	47	51	+4	46	44	−2
4	41	42	+1	43	40	−3
5	40	41	+1	48	50	+2
6	52	55	+3	47	45	−2
7	52	59	+7	51	50	−1
8	40	41	+1	45	48	+3
9	25	25	0	37	27	−10
10	40	45	+5	40	43	+3
11	43	42	−1	52	42	−10
12	59	55	−4	62	54	−8
13	42	39	−3	45	38	−7
14	28	27	−1	41	28	−13
15	51	50	−1	53	37	−16
Average	42·6	43·3	+0·7	46·8	41·5	−5·3
Average difference, neglecting signs		2·3			6·4	

18. We shall deal with a comparison of the two sets of numerical marks later. In some ways it is more interesting to compare the variation of the examiners' judgments in regard to the awards of Failure, Pass, and Credit. They are set out in Table 11 below :—

TABLE 11

AWARDS OF FAILURE, PASS, AND CREDIT (F, P AND C)

	Examiner													
	B		C		D		E		F		G		H	
Candidate	1st Marking	2nd Marking	1st Marking	2nd Marking	1st Marking	2nd Marking	1st Marking	2nd Marking	1st Marking	2nd Marking	1st Marking	2nd Marking	1st Marking	2nd Marking
1	F	F	P	P	P	F	F	C	F	P	P	P	P	C
2	F	F	F	P	P	F	P	P	F	F	F	P	P	F
3	F	F	C	C	P	F	P	F	P	F	P	C	P	P
4	P	P	C	C	P	P	C	C	C	C	C	P	C	P
5	F	F	P	C	P	F	P	P	P	P	P	C	P	C
6	F	C	F	F	C	F	C	P	C	C	C	C	C	C
7	P	P	C	C	P	P	C	C	C	P	C	C	C	C
8	F	F	P	F	F	F	F	P	P	F	P	F	P	F
9	F	F	F	F	F	F	F	F	F	F	P	F	P	P
10	P	P	F	C	P	F	F	P	P	F	C	P	C	P
11	F	F	C	P	P	F	P	F	C	P	P	P	P	C
12	P	P	C	C	P	F	C	C	C	P	C	C	C	C
13	F	F	F	F	F	F	F	F	F	F	F	P	P	P
14	P	P	F	P	P	F	P	F	P	F	C	C	C	P
15	P	C	C	C	C	C	P	C	C	C	C	C	C	C
No change	13		9		6		7		7		7		7	
No. of candidates moved one class up	1		3				3		1		4		3	
No. of candidates moved one class down			2		8		4		7		4		5	
No. of candidates moved two classes up	1		1				1							
No. of candidates moved two classes down					1									
Total No. of Failures	9	8	6	4	3	12	5	5	4	7	2	2	0	2
Total No. of Passes	6	5	3	4	10	2	6	5	5	5	6	6	8	6
Total No. of Credits	0	2	6	7	2	1	4	5	6	3	7	7	7	7

TABLE 11—*continued*

Candidate	Examiner J 1st Marking	J 2nd Marking	K 1st Marking	K 2nd Marking	L 1st Marking	L 2nd Marking	M 1st Marking	M 2nd Marking	N 1st Marking	N 2nd Marking	P 1st Marking	P 2nd Marking	Q 1st Marking	Q 2nd Marking
1	F	C	C	P	P	P	P	P	F	P	P	P	C	P
2	F	F	F	F	F	F	P	P	F	F	F	F	P	F
3	P	P	C	C	F	P	P	F	F	F	P	C	P	P
4	C	F	C	P	F	F	C	P	C	C	P	P	P	P
5	P	P	C	P	P	P	F	P	F	P	P	P	P	C
6	P	C	P	P	F	P	P	C	P	P	C	C	P	P
7	C	C	C	P	P	P	P	C	C	P	C	C	C	C
8	F	F	P	F	F	F	P	F	F	F	P	P	P	P
9	P	F	P	F	F	F	P	F	F	F	F	F	F	F
10	C	F	C	C	P	P	P	P	P	P	P	P	P	P
11	C	F	C	P	F	F	P	P	F	P	P	P	C	P
12	C	C	C	C	P	F	C	P	P	C	C	C	C	C
13	F	F	P	P	F	F	P	P	F	P	P	F	P	F
14	P	P	P	P	P	F	P	P	F	P	F	F	P	F
15	C	C	C	C	C	P	C	C	C	P	C	C	C	F
No change	9		8		10		7		7		13		8	
No. of candidates moved one class up	1				2		3		6		1		1	
No. of candidates moved one class down	1		7		3		5		2		1		5	
No. of candidates moved two classes up	1													
No. of candidates moved two classes down	3												1	
Total No. of Failures	4	7	1	3	8	8	1	3	9	4	3	4	1	5
Total No. of Passes	5	3	5	8	6	7	11	9	3	9	8	6	9	7
Total No. of Credits	6	5	9	4	1	0	3	3	3	2	4	5	5	3

19. It will be seen that when the fourteen examiners came to make the awards of Failure, Pass, and Credit to the fifteen several candidates on the second occasion, they had changed their minds in 92 cases out of a total of 210.

20. The examiners were on the whole more severe on the second occasion than the first. The number of Failures, Passes, and Credits given by the individual examiners, B to Q, on the two occasions are shown at the foot of Table 11.

21. Examiners B and P are the most consistent in their awards. Examiner J keeps nine candidates in their original classes, but moves one candidate two classes up and three two classes down (from Credit to Failure). Examiner D moves eight candidates one class down and one two classes down. Examiner K moves seven candidates down one class. G moves four candidates up one class and four candidates down one class.

The case of G and similar cases are particularly interesting because they show that an examiner may return the same number of Failures, Passes and Credits on two occasions, but place different candidates in these categories. Steadiness in statistics of classification is therefore not necessarily a proof of steadiness of judgment. The ordinary methods of scrutinising the marks of an assistant-examiner after his first trial scripts have been inspected would completely fail to reveal variability of this kind. J's classification statistics are not very different on the two occasions (he diminishes the number of Credits by one and of Passes by two), but he alters the class of six candidates out of fifteen.

Only B and P make substantially the same awards on the two occasions.

22. We now turn to the numerical results. The marks in Table 9 have been adjusted, in view of the difference of the limits fixed by the individual examiners for Pass and Credit noted at the bottom of Table 9, in exactly the same way as the marks in Table 1 were adjusted (see para. 7 above) to render them more comparable.[1]

It may be pointed out that six examiners, B, E, J, K, P and Q, kept their limits unaltered, while some of the rest altered the limiting marks by as much as 5 or 6 units, and the difference between those limits by as much as 4 units.

23. Table 10 above shows the adjusted marks of both the first and second markings of each examiner, with the differences between the two markings, the average marks of the two markings, the average differences taking signs into account, and the average differences neglecting signs.

24. The differences shown earlier in the classification of candidates are in some cases more strikingly evident in the numerical marks. Thus Examiner E gives Candidate No. 1 30 marks on the first occasion and 55 on the second; in two other cases he gives 10 or 11 marks more on the second occasion than on the first, while in two others he gives 10 or 11 marks less.

[1] It will be remembered that this adjustment does not alter the order of the candidates or their classification.

His average only varies by a unit and he awards the same number of Failures, one less Pass, one more Credit. Yet he changes his mind in regard to the fate of eight candidates out of fifteen. One is bound to ask oneself whether this irregularity of judgment would not display itself equally during the progress of marking 500 papers on the same occasion. It is, for the candidates concerned, a formidable irregularity; yet it is important to note that it is one which no head-examiner could possibly detect by the ordinary methods employed. E, with this remarkable unsteadiness of judgment in regard to the performance of individual candidates, would in all probability be regarded as a steady examiner just because he kept his statistics steady, unless some flagrant case of mis-marking on the border-line attracted attention.

25. In some cases the examiner has altered his general standard on the second occasion. D gives lower marks than on the first throughout, and moves eight candidates down a class (and one two classes), and K gives lower marks throughout and moves seven candidates down a class. Such changes might possibly be detected in a real examination; but if they occurred on one day and were compensated for on the next in the course of marking 500 scripts during, say, a week's work or a fortnight's work, the chance of detection would be small. In an examination where school-records were consulted, some of the irregularities might be detected. But it is difficult to see at present how in a large-scale examination such irregularities of judgment could be adequately dealt with in the interest of all the candidates.

26. Of the fourteen examiners there is one, P, who is exceptionally steady and whose mark never varies by more than 7; B, though he gives one difference of 14, and another of 13 marks, is also much above the average in steadiness (see Table 10 above).

27. It may perhaps be noted here that four of the examiners (B, E, F, P) volunteered the opinion that the question-paper was of average standard, and that five (C, D, H, K, Q) regarded it, for various reasons, as difficult. The marks of those examiners who thought the paper difficult were on the whole higher than the marks of those who thought that the paper was of the ordinary standard. The average of the averages of B, E, F, P was 41·9 on the first, and 41·3 on the second occasion; those of C, D, H, K, Q were 47·0 on the first and 43·3 on the second occasion. Only one examiner, K, said (on both occasions) that he had marked leniently owing to the difficulty of the paper.

CHAPTER II

MARKING OF SCHOOL CERTIFICATE LATIN SCRIPTS

28. *Character of the Examination Papers.*—The scripts investigated were written in answer to two two-hour papers. We give below a summary of the questions in each paper together with the maximum marks assigned to each.

PAPER I.

Qns. 1, 2 and 3 were on accidence. Maximum for each Qn., 10 marks.

Qn. 4. Short passage for translation, with explanation. Maximum, 10 marks.

Qn. 5. English sentences for translation into Latin. Maximum, 50 marks.

Qn. 6. (An alternative to Qn. 5.) An English passage for translation into Latin. Maximum, 50 marks.

Qn. 7. Two Latin passages (unseen), one in prose and one in verse, for translation into English. Maximum, 60 marks.

Total marks for the paper, 150.

PAPER II (Set Books).

Qn. 8. Two passages of prose to be translated into English, 30 marks each. Maximum, 60 marks.

Qn. 9. Short passages of prose for translation and comment. Maximum, 10 marks.

Qn. 10. A request for explanatory notes on certain passages in prose. Maximum, 10 marks.

Qn. 11. An historical question. Maximum, 16 marks.

Qn. 12. Two passages of verse for translation into English. Maximum, 60 marks.

Qn. 13. A question on scansion. Maximum, 8 marks.

Qn. 14. Passages of verse for translation and grammatical comment. Maximum, 10 marks.

Qn. 15. Two short passages in verse for explanatory comment. Maximum, 10 marks.

Qn. 16. A question on the narrative of the poem. Maximum, 16 marks.

Total marks for the paper, 200.

29. *Special Object of the Investigation and Method of Selection of Scripts.*—The object of the investigation was similar in some ways to that of the investigation on School Certificate History Scripts, namely, to examine the discrepancies between the markings of scripts which on the first marking appeared to be of approximately the same value. It was, however, in some ways more complex than the investigation on the History scripts, since the scripts were written as answers to two examination-papers instead of one.

The scripts of fifteen candidates were supplied by the School Certificate authority concerned, thirty in all, so selected that the candidates had obtained exactly the same moderate mark for the two papers combined. As in other cases, every trace of the origin of the scripts and of the original marking was removed from the scripts before they were circulated to the examiners. No marks were made on the scripts by the examiners.

30. *Procedure.*—Fifteen examiners were selected from the panel of a School Certificate authority, other than the one by whom the scripts had been furnished. The examiners were informed that the object of the investigation was to form an estimate of the differences in marking between different examiners, but as in the case of History they were not informed that the scripts had been originally marked as being of the same value.

31. The scripts were first submitted to two examiners of the panel, who were treated as Chief Examiners, and who, after they had marked the scripts independently on the basis of the original marking-scheme, were asked to be responsible for a revised marking-scheme.

Simultaneously, copies of the examination-papers (not the scripts) together with the original marking-scheme were sent to the other thirteen examiners for comment, and their comments were then considered by the two Chief Examiners.

As a result of this procedure, *two* final marking-schemes, more detailed than the original marking-scheme, were framed by the Chief Examiners, which differed between themselves in only one respect, viz., that more detailed instructions were given for unprepared passages for translation from and into Latin under Scheme I than under Scheme II, the maximum for each question remaining the same.

32. The Chief Examiners drew up fourteen additional general instructions included in both Scheme I and Scheme II, together with a sample passage marked by them. Scheme I included nineteen more detailed instructions than Scheme II in respect of the unprepared passages for translation from and into Latin in Paper I, Qns. 6 and 7.[1] There was no other difference between the two schemes.

33. Scheme I was supplied to the six examiners A to F, designated below as Group I, who had stated that they preferred the more detailed instructions; and Scheme II was supplied to the seven examiners, G to N, designated below as Group II, who preferred the less detailed instructions.

34. In the statistical analysis we have considered the results of the two Groups both separately and together. It is obvious that these two Groups cannot be strictly regarded as analogous to two independent Boards, who would no doubt have adopted marking-schemes differing far more widely.

In order to facilitate marking in considerable detail, the original maximum for Paper I was, as shown above, 150, and the original maximum for Paper II was 200. The two papers were, however, treated both by the original authority and in this investigation as of equal value; and the marks allotted for Paper I in the first instance were therefore divided by 3 and those for Paper II were divided by 4, so as to yield a maximum of 50 marks for each paper, i.e., 100 for the two papers.

35. In Table 12 below, we give the *total* marks awarded on this basis by the thirteen examiners to the fifteen candidates.

[1] Of these, 10 of the detailed instructions were given in respect of Qn. 6, which was attempted only by a single candidate (see para. 40 below).

TABLE 12

MARKS ALLOTTED TO THE FIFTEEN COUPLES OF SCRIPTS

Two Papers. (*Maximum for the two papers* = 100)

Cand.	*Group I* Examiner						Range	*Group II* Examiner							Range
	A	B	C	D	E	F		G	H	J	K	L	M	N	
1	39	43	52	37	43	40	15	41	44	52	41	52	43	37	15
2	39	44	50	43	43	46	11	38	43	54	44	52	44	41	16
3	44	51	55	47	46	46	11	45	49	61	42	60	53	50	19
4	37	46	43	44	40	43	9	39	40	55	38	46	44	43	17
5	38	47	55	35	43	45	20	49	42	58	40	51	42	49	18
6	45	50	54	45	45	49	9	52	47	61	44	59	49	51	17
7	42	52	51	45	44	46	10	49	46	53	49	56	49	51	10
8	43	49	53	47	46	46	10	47	49	52	47	51	48	48	5
9	32	42	49	34	36	38	17	42	37	53	36	46	41	40	17
10	37	40	48	37	39	42	11	40	38	52	35	44	38	45	17
11	38	42	47	39	36	39	11	50	40	55	40	49	42	42	15
12	40	44	50	41	36	42	14	43	42	52	39	48	42	44	13
13	38	43	50	36	34	41	16	41	40	52	38	50	37	43	15
14	35	45	49	37	40	40	14	46	42	57	37	55	42	44	20
15	32	38	41	28	34	34	13	33	33	50	33	45	38	41	17
Averages	38·6	45·1	49·8	39·7	40·3	42·5	12·7	43·7	42·1	54·5	40·2	50·9	43·5	44·6	15·4
	Average of Average Marks = 42·7							Average of Average Marks = 45·6							

36. It is to be pointed out at once that whereas the fifteen couples of scripts had all originally been assigned the same moderate mark, under Scheme I they received 24 different marks, ranging from 28 to 55, and under Scheme II, 28 different marks, ranging from 33 to 61.[1]

In this investigation the examiners were not asked to allot awards of Pass, Credit, and Distinction, as they were in other investigations. We have only considered numerical marks.

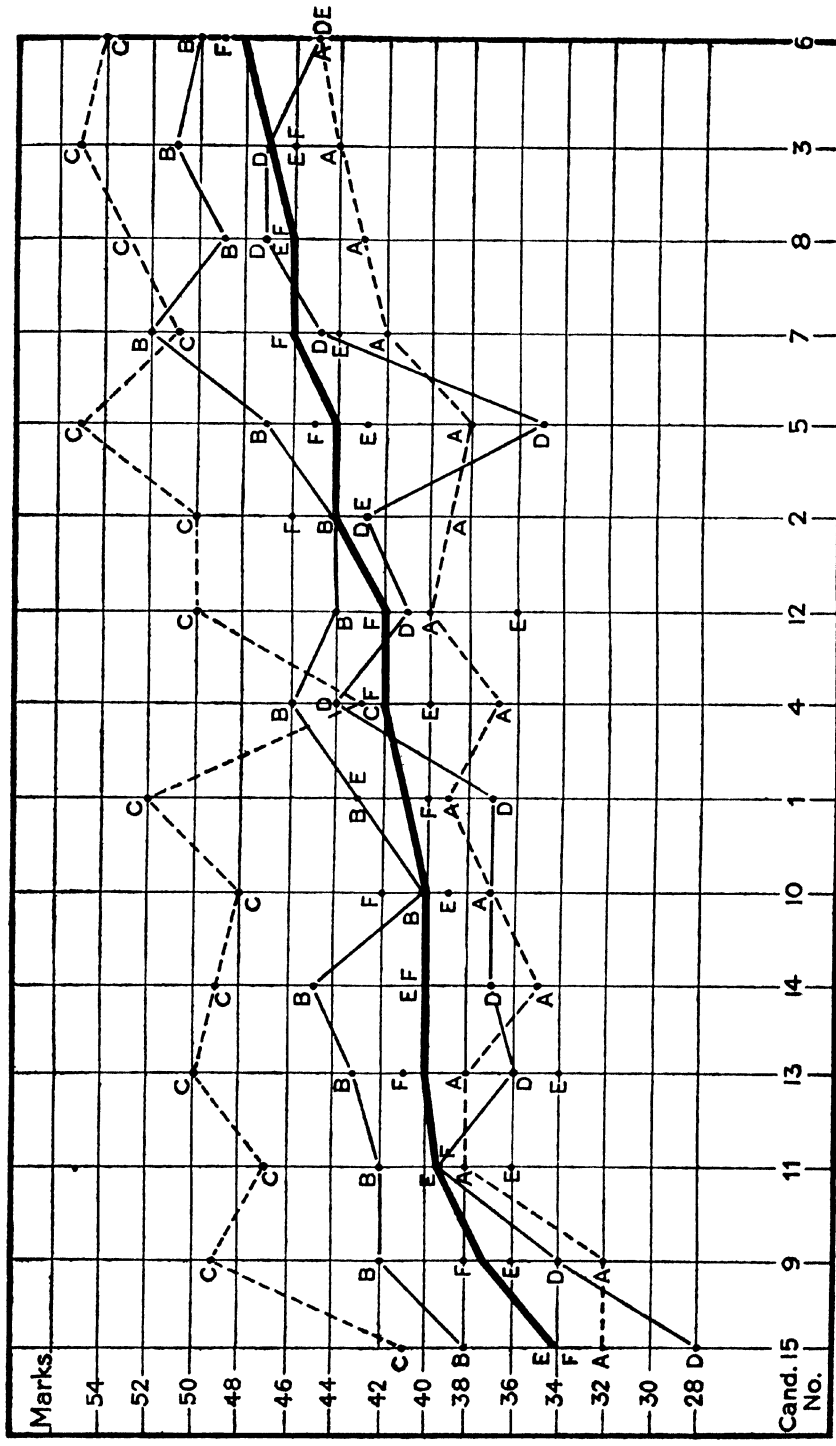
37. If we now consider the marks allotted to the individual candidates by the six examiners of Group I, we find that the range (i.e. the difference between the highest and lowest mark allotted to a candidate) varies from 9 to 20, with an average of 12·7; and that with the seven examiners of Group II (who worked with the less detailed scheme of marking) the range varies from 5 to 20, with an average of 15·4. The marks of Examiner J are exceptionally high. His average is 54·5, whereas the average of the averages for his group is only 45·6; again, the average of Examiner C is 49·8 as against the average of averages of 42·6.

38. The marks of individual examiners are set out in diagrammatic form on the following pages. The order of the candidates is the order in ascending magnitude of the "Ideal"[2] marks

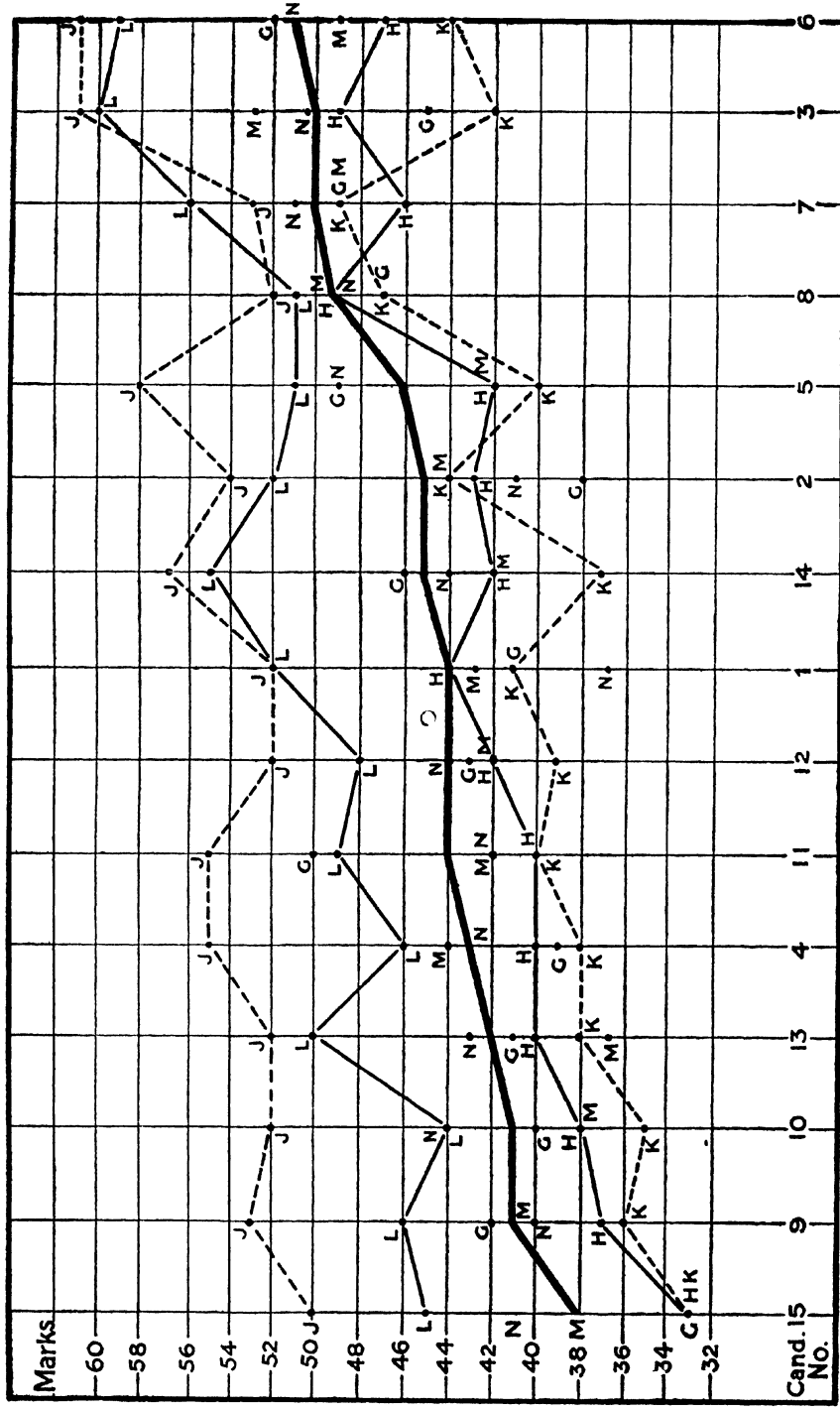
[1] The total number of different marks allotted under the two Schemes was 31, and the total range, from 28 to 61.

[2] The term "Ideal Mark" is explained in Part II (p. 186 below).

SCHOOL CERTIFICATE LATIN—GROUP I



SCHOOL CERTIFICATE LATIN—GROUP II



allotted to them in accordance with the method of calculation explained in Part II. The Ideal marks are indicated by the thick lines in the middle of the two diagrams. The interest of the diagram lies in the fact that certain examiners are almost always above or below their colleagues. For instance, the graphs for C and A in Group I never intersect, and similarly in Group II the graphs for J and K. In order not to overcrowd the diagrams, the graphs for Examiners E and F in Group I and for Examiners G, M and N in Group II are omitted. The actual marks awarded by all these examiners are given in Table 12. In an ordinary examination the marks of the exceptional examiners might have been scaled down to a figure nearer the average. But in the present investigation we wished to examine the variations possible when experienced examiners in Latin are working on the same detailed scheme but independently.

39. We shall deal more in detail later with the idiosyncrasies of the different examiners.

We now turn to the marks for answers to individual questions, for the comprehension of which a fuller description of the marking-scheme is necessary.

40. We give in Tables 13, 14 and 15 below a general analysis of the marks allotted to the answers by the two Groups. But we may note that Group I had exactly the same instructions as Group II except in respect of Qn. 7, on which we shall comment specially. We have omitted from the Table the figures for Qn. 6 as it was answered by only a single candidate.

41. We see from Table 15 that candidates earn the highest average marks for Qns. 1 and 2 (accidence), and for Qn. 8 (prose set books). The average marks for the rest of the paper are many of them considerably lower. The highest average per cent. is 76·7 for an easy question in accidence, and the lowest 15·3 for Qn. 7, translation from Latin unseen into English.

The lowest average range is for Qn. 1 (accidence), the highest for Qn. 15 (explanatory comment on two short passages), in which the difference of standards between different examiners is very high.

Generally speaking there is a sharp difference between the questions on accidence, in which the concurrence of examiners is (as was to be expected) very high, and those on translation, grammatical comment, and history, in which the examiners' marks were wide apart.

42. As will be seen from Tables 13 to 15, the examiners of Group II on the whole mark much more generously than those of Group I. The greatest differences between the averages of the averages for the two Groups are for Qns. 9, 10 and 14, all

TABLE 13 GROUP I

	Question	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16
(a)																
(b)	Number of Candidates	15	15	15	15	14	14	15	14	13	14	15	12	11	14	14
(c)	Maximum Marks	10	10	10	10	50	60	60	10	10	16	60	8	10	10	16
(d)	Average Marks — Exmr. A	6·47	7·60	4·60	5·27	12·21	6·00	44·13	3·36	2·15	8·71	27·47	2·83	0·91	4·14	7·93
	Exmr. B	6·47	7·67	5·00	5·27	16·21	13·43	43·07	4·00	2·85	9·00	36·13	3·17	2·09	4·36	8·00
	Exmr. C	6·47	7·60	5·00	6·00	17·36	9·57	50·00	3·36	2·46	12·64	40·73	3·58	2·27	8·21	10·86
	Exmr. D	6·53	7·53	5·07	5·20	15·43	5·93	40·67	2·64	2·46	7·71	29·20	2·83	2·54	4·71	8·57
	Exmr. E	6·47	7·67	4·57	5·33	11·79	9·51	40·93	2·61	1·15	8·57	32·73	3·17	1·23	3·68	9·64
	Exmr. F	6·47	7·67	4·13	5·20	15·43	10·50	44·93	3·50	2·15	7·50	32·93	3·25	1·64	6·07	6·00
(e)	Averages of Examiners' Averages	6·48	7·62	4·73	5·38	14·74	9·16	43·95	3·24	2·20	9·02	33·20	3·14	1·78	5·19	8·50
(f)	% of Maximum Mark	64·8	76·2	47·3	53·8	29·5	15·3	73·3	32·4	22·0	56·4	55·3	39·2	17·8	51·9	53·1
(g)	Highest Average Marks	6·53	7·67	5·07	6·00	17·36	13·43	50·00	4·00	2·85	12·64	40·73	3·58	2·54	8·21	10·86
	Lowest Average Marks	6·47	7·53	4·13	5·20	11·79	5·93	40·67	2·61	1·15	7·50	27·47	2·83	0·91	3·68	6·00
(h)	Difference	0·06	0·14	0·94	0·80	5·57	7·50	9·33	1·39	1·70	5·14	13·26	0·75	1·63	4·53	4·86
(j)	% of Maximum Mark	0·6	1·4	9·4	8·0	11·1	12·5	15·5	13·9	17·0	32·1	22·1	9·4	16·3	45·3	30·4
(k)	Average Range	0·20	0·33	1·13	1·40	7·43	9·43	12·47	2·54	2·62	6·79	15·80	1·50	2·18	5·32	6·21
(l)	% of Maximum Mark	2·0	3·3	11·3	14·0	14·9	15·7	20·8	25·4	26·2	42·4	26·3	18·7	21·8	53·2	38·8
(m)	Order of Examiners (Average Marks) A	—	—	4	3½	5	5	3	3½	4½	3	6	5½	6	5	5
	B	—	—	2½	3½	2	1	4	1	1	2	2	3½	3	4	4
	C	—	—	2½	1	1	3	1	3½	2½	1	1	1	2	1	1
	D	—	—	1	5½	3½	6	6	5	2½	5	5	5½	1	3	3
	E	—	—	5	2	6	4	5	6	6	4	4	3½	5	6	2
	F	—	—	6	5½	3½	2	2	2	4½	6	3	2	4	2	6

TABLE 14 GROUP II

(a)	Question		1	2	3	4	5	7	8	9	10	11	12	13	14	15	16
(b)	Number of Candidates		15	15	15	15	14	14	15	14	13	14	15	12	11	14	14
(c)	Maximum Marks		10	10	10	10	50	60	60	10	10	16	60	8	10	10	16
(d)	Average Marks	Exmr. G	6·67	7·67	5·07	5·80	15·57	14·50	35·47	4·00	2·85	9·71	31·60	5·00	3·18	5·79	8·57
		,, H	6·47	7·67	5·00	5·13	15·64	4·21	42·07	4·14	3·23	10·00	35·40	3·42	2·27	4·64	8·86
		,, J	6·47	7·87	5·00	5·20	19·14	22·00	51·47	4·36	3·46	9·21	47·53	4·08	3·00	4·79	7·36
		,, K	6·60	7·60	4·87	5·80	13·71	6·36	40·67	3·50	2·31	7·00	30·87	3·50	3·54	4·29	9·86
		,, L	6·47	7·47	4·80	5·90	20·04	15·71	47·80	4·07	2·92	12·29	43·00	2·92	2·54	3·57	9·43
		,, M	6·47	7·73	5·00	5·00	14·00	11·07	43·53	4·00	2·46	9·86	33·27	3·58	2·36	5·29	9·00
		,, N	6·47	7·67	4·73	5·40	15·29	14·00	43·53	3·64	3·15	9·93	33·93	3·25	3·18	4·29	7·57
(e)	Averages of Examiners' Averages		6·52	7·67	4·92	5·46	16·20	12·55	43·51	3·96	2·91	9·71	36·51	3·68	2·87	4·67	8·66
(f)	% of Maximum Mark		65·2	76·7	49·2	54·6	32·4	20·9	72·5	39·6	29·1	60·7	60·9	46·0	28·7	46·7	54·1
(g)	Highest Average Marks		6·67	7·87	5·07	5·90	20·04	22·00	51·47	4·36	3·46	12·29	47·53	5·00	3·54	5·79	9·86
	Lowest Average Marks		6·47	7·47	4·73	5·00	13·71	4·21	35·47	3·50	2·31	7·00	30·87	2·92	2·27	3·57	7·36
(h)	Difference		0·20	0·40	0·34	0·90	6·33	17·79	16·00	0·86	1·15	5·29	16·66	2·08	1·27	2·22	2·50
(j)	% of Maximum Mark		2·0	4·0	3·4	9·0	12·7	29·6	26·7	8·6	11·5	33·1	27·8	26·0	12·7	22·2	15·6
(k)	Average Range		0·47	0·67	0·67	1·53	8·89	18·64	16·47	3·07	3·00	6·00	19·93	2·67	2·91	3·86	5·21
(l)	% of Maximum Mark		4·7	6·7	6·7	15·3	17·8	31·1	27·4	30·7	30·0	37·5	33·2	33·4	29·1	38·6	32·6
(m)	Order of Examiners (Average Marks)	G	—	—	1	2½	4	3	7	4½	5	5	6	1	2½	1	5
		H	—	—	3	6	3	7	5	2	2	2	3	5	7	4	4
		J	—	—	3	5	2	1	1	1	1	6	1	2	4	3	7
		K	—	—	5	2½	7	6	6	7	7	7	7	4	1	5½	1
		L	—	—	6	1	1	2	2	3	4	1	2	7	5	7	2
		M	—	—	3	7	6	5	3½	4½	6	4	5	3	6	2	3
		N	—	—	7	4	5	4	3½	6	3	3	4	6	2½	5½	6

TABLE 14A GROUP I AND GROUP II COMBINED

(a)	Question	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16
(b)	Number of Candidates	15	15	15	15	14	14	15	14	13	14	15	12	11	14	14
(c)	Maximum Marks	10	10	10	10	50	60	60	10	10	16	60	8	10	10	16
(d)	Average Range	0·60	0·80	1·33	1·97	10·32	—	18·67	3·86	3·62	7·64	24·40	2·92	3·91	5·93	7·07
(e)	% of Maximum Mark	6·0	8·0	13·3	19·7	20·6	—	31·1	38·6	36·2	47·7	40·7	36·5	39·1	59·3	44·2

TABLE 15

COMPARISON OF MARKS OF THE TWO GROUPS

(The marks are expressed as percentages of the maximum in each case)

(a)	Question	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16
	Averages of Examiners' Averages															
(b)	Group I	64·8	76·2	47·3	53·8	29·5	15·3	73·3	32·4	22·0	56·4	55·3	39·2	17·8	51·9	53·1
	Group II	65·2	76·7	49·2	54·6	32·4	20·9	72·5	39·6	29·1	60·7	60·9	46·0	28·7	46·7	54·1
	Difference	−0·4	−0·5	−1·9	−0·8	−2·9	−5·6	+0·8	−7·2	−7·1	−4·3	−5·6	−6·8	−10·9	+5·2	−1·0
	Difference between Highest and Lowest Averages															
(c)	Group I	0·6	1·4	9·4	8·0	11·1	12·5	15·5	13·9	17·0	32·1	22·1	9·4	16·3	45·3	30·4
	Group II	2·0	4·0	3·4	9·0	12·7	29·6	26·7	8·6	11·5	33·1	27·8	26·0	12·7	22·2	15·6
	Difference	−1·4	−2·6	+6·0	−1·0	−1·6	−17·1	−11·2	+5·3	+5·5	−1·0	−5·7	−16·6	+3·6	+23·1	+14·8
	Average Range															
(d)	Group I	2·0	3·3	11·3	14·0	14·9	15·7	20·8	25·4	26·2	42·4	26·3	18·7	21·8	53·2	38·8
	Group II	4·7	6·7	6·7	15·3	17·8	31·1	27·4	30·7	30·0	37·5	33·2	33·4	29·1	38·6	32·6
	Difference	−2·7	−3·4	+4·6	−1·3	−2·9	−15·4	−6·6	−5·3	−3·8	+4·9	−6·9	−14·7	−7·3	+14·6	+6·2

of which are over 7%; and the lowest for Qns. 1, 2, 4 and 8, which are all under 1%.

The two Tables 13 and 14 show that for Qn. 14 the averages of the examiners of Group I range from 0·91 to 2·54 marks, whereas those of Group II vary from 2·27 to 3·54, the maximum mark being 10.

43. A contrast of the differences of standard between the several examiners of Group I and between the several examiners of Group II is interesting.

From Table 13 we see that in regard to Qn. 15 the several examiners of Group I have very different standards, since the difference between the highest and the lowest average (see rows *g*, *h* and *j*) is 45·3% of the maximum mark (10). The corresponding figure in Table 14 (row *j*) for Qn. 15 is only 22·2% of the maximum mark.

On the other hand, in Qn. 13 several examiners of Group I agree fairly well with one another, the difference between the highest and the lowest average being only 9·4% of the maximum mark (Table 13, row *j*), whereas the corresponding figure for Group II is 26% of the maximum mark (see Table 14, row *j*).

On the whole, the examiners of Group II show greater differences of standard between themselves than those of Group I; and this fact translates itself in another way in the fact that Group II have higher average ranges than Group I (see rows *k* and *l* of Tables 13 and 14).

44. Examiners show some consistency in marking higher or lower than the majority of their colleagues; thus, in Group I, Examiner C has a high average mark in all the questions, while Examiner A has a low one (see Table 13, row *d*).

Similarly, in Group II, Examiners J and L have relatively high average marks for most questions, while Examiner K has a low one (see Table 14, row *d*).

45. The immediately preceding paragraphs deal only with questions of which the maximum is low. We extract below details of questions for which the maximum is either 50 or 60 marks, and which are therefore of great importance from the point of view of the candidates. The figures given are expressed as percentages of the maximum mark in each case.

		Question				
		5	7	8	12	
	Maximum	50	60	60	60	
Average of Averages	Group I	29·5	15·3	73·3	55·3	Table 13, row *f*
	Group II	32·4	20·9	72·5	60·9	Table 14, row *f*
	Difference	−2.9	−5·6	+0·8	−5·6	

		Question 5	7	8	12	
Difference between highest and lowest averages	Maximum	50	60	60	60	
	Group I	11·1	12·5	15·5	22·1	Table 13, row *j*
	Group II	12·7	29·6	26·7	27·8	Table 14, row *j*
	Difference	−1·6	−17·1	−11·2	−5·7	
Average range	Group I	14·9	15·7	20·8	26·3	Table 13, row *l*
	Group II	17·8	31·1	27·4	33·2	Table 14, row *l*
	Difference	−2·9	−15·4	−6·6	−6·9	

46. It will be seen that candidates score many marks on Qn. 8 (translation of prose set books) and Qn. 12 (translation of verse set book), but do not score many on Qn. 5 (translation of English sentences into Latin) or on Qn. 7 (translation from Latin unseen). The differences between the standards of the two Groups in respect of these four questions are not great; but, on the other hand, the differences between the highest and lowest averages of Group II for Qns. 7, 8 and 12 are much greater than in the case of examiners of Group I, and are seriously disturbing. Even for Group I the ranges are high for Qns. 8 and 12 and show serious differences between the standards of the different examiners.

47. We give below in Table 16, as an example of close marking, the marks for Qn. 1 :—

TABLE 16

LATIN I. *Question* 1 (*Accidence*) *Maximum* = 10 *marks*

Cand.	*Group I* Examiner A	B	C	D	E	F	Range	*Group II* Examiner G	H	J	K	L	M	N	Range	Extreme Range
1	6	6	6	6	6	6	0	6	6	6	6	6	6	6	0	0
2	5	5	5	6	5	5	1	5	5	5	5	5	5	5	0	1
3	6	6	6	6	6	6	0	6	6	6	6	6	6	6	0	0
4	9	9	9	9	9	9	0	9	9	9	9	9	9	9	0	0
5	6	6	6	6	6	6	0	6	6	6	6	6	6	6	0	0
6	7	7	7	6	7	7	1	8	7	7	7	7	7	7	1	2
7	7	7	7	8	7	7	1	8	7	7	7	7	7	7	1	1
8	7	7	7	7	7	7	0	7	7	7	7	7	7	7	0	0
9	4	4	4	4	4	4	0	4	4	4	5	4	4	4	1	1
10	6	6	6	6	6	6	0	5	6	6	6	6	6	6	1	1
11	7	7	7	7	7	7	0	7	7	7	7	7	7	7	0	0
12	6	6	6	6	6	6	0	7	6	6	6	6	6	6	1	1
13	8	8	8	8	8	8	0	8	8	8	9	8	8	8	1	1
14	7	7	7	7	7	7	0	8	7	7	7	7	7	7	1	1
15	6	6	6	6	6	6	0	6	6	6	6	6	6	6	0	0
Averages	6·5	6·5	6·5	6·5	6·5	6·5	0·2	6·7	6·5	6·5	6·6	6·5	6·5	6·5	0·5	0·6

48. Here the differences between examiners are insignificant. A detailed analysis might show that they depend on different interpretations of not too legible handwritings.

49. As an intermediate type we might cite Qn. 4 (Latin sentences to be translated and explained), in which the average range (on a maximum of 10) is only 1·4 marks for Group I and 1·5 marks for Group II (while the average of the extreme ranges for the two groups is 2 marks).

50. We come now to another kind of question with a maximum of 10 marks, Qn. 15, which demanded explanatory notes on the subject of two short passages in verse. The marks are given in full in Table 17 below :—

TABLE 17

(*Directions the same for both Groups*)

LATIN II. *Qn.* 15 (*Explanatory notes on subject-matter of two short passages in verse*) *Maximum* = 10 *marks*

Candidate	*Group I* Examiner						Range	*Group II* Examiner							Range	Extreme Range
	A	B	C	D	E	F		G	H	J	K	L	M	N		
1	5	7	9	4	6	9	5	7	6	9	5	10	6	5	5	6
3	6	7	10	4	6	9	6	10	8	6	5	8	8	6	5	6
4	1	2	4	1	3	4	3	3	3	2	1	2	3	2	2	3
5	7	6	10	4	5	7	6	7	5	3	5	3	8	6	5	7
6	8	6	10	10	5	8	5	9	7	8	7	4	8	7	5	6
7	4	5	9	8	4	7	5	9	5	6	6	4	7	5	5	5
8	1	3	9	1	1	5	8	7	2	6	2	1	4	3	6	8
9	4	4	8	5	4	6	4	6	5	4	6	2	4	3	4	6
10	5	4	9	7	3½	5	5½	5	6	5	4	2	5	6	4	7
11	3	3	5	2	2	4	3	4	4	2	3	5	5	2	3	3
12	4	4	10	7	3	6	7	5	5	6	5	2	6	5	4	8
13	1	2	10	3	1	3	9	2	2	2	2	1	2	2	1	9
14	4	3	6	2	4	5	4	4	3	4	3	2	3	3	2	4
15	5	5	6	8	4	7	4	3	4	4	6	4	5	5	3	5
Averages	4·1	4·4	8·2	4·7	3·7	6·1	5·3	5·8	4·6	4·8	4·3	3·6	5·3	4·3	3·9	5·9

51. Here, as in the previous questions, the two Groups had exactly the same instructions ; the average range for Group I was 5·3, for Group II, 3·9 ; for the two Groups together, 5·9. These are very large differences for a question with a maximum of 10 marks—an average extreme range of nearly 60% of the maximum. The Table shows that the variation in the marking of short questions of the essay type is of the same order as the variations in the marking of longer questions of this type and of translations. The discrepancies for this particular question are higher indeed than for any other question. But remembering

that the maximum is only 10 and that (as stated in para. 34 above) the marks were divided by 4 before being reckoned as part of the final total maximum of 100, the differences between examiners, though so large in their first expression, become almost negligible so far as the fate of individual candidates is concerned.

52. When we come to the questions on translation to and from Latin, to which maxima of 50 and 60 were allotted in the first instance, and on which candidates earned a comparatively large proportion of their marks, it is a different matter. We shall give three examples of the marks for this type of question. Table 18 below gives the marks for Qn. 8, a piece of translation from a Latin prose author (set book) with a maximum of 60 marks.

TABLE 18

(*Directions the same for both Groups*)

LATIN II. *Qn.* 8 (*Two passages from Cæsar for translation into English*)
Maximum = 60 *marks*

Cand.	*Group I* Examiner						Range	*Group II* Examiner							Range	Extreme Range
	A	B	C	D	E	F		G	H	J	K	L	M	N		
1	44	44	49	40	41	44	9	33	43	50	38	46	45	35	17	17
2	46	41	52	51	44	48	11	28	43	53	41	55	44	45	27	27
3	50	46	58	50	48	48	12	41	51	54	45	52	53	52	13	17
4	50	45	50	53	44	48	9	34	48	51	44	49	48	49	17	19
5	51	46	51	50	43	48	8	39	43	54	43	54	42	54	15	15
6	48	43	51	46	39	48	12	36	44	51	38	44	44	46	15	15
7	56	57	57	51	50	52	7	43	50	57	59	56	50	53	16	16
8	49	52	53	50	53	53	4	46	50	55	51	54	53	52	9	9
9	33	38	45	31	33	40	14	23	33	47	34	42	41	35	24	24
10	37	32	42	17	36	43	26	31	33	45	31	40	36	35	14	28
11	48	49	53	48	44	50	9	46	45	54	43	54	49	45	11	11
12	48	44	47	33	40	45	15	37	42	53	41	46	44	44	16	20
13	26	35	46	23	28	34	23	27	32	47	28	38	30	35	20	24
14	43	44	53	43	43	43	10	39	45	54	44	52	44	41	15	15
15	33	30	43	24	28	30	19	29	29	47	30	35	30	32	18	23
Averages	44·1	43·1	50·0	40·7	40·9	44·9	12·5	35·5	42·1	51·5	40·7	47·8	43·5	43·5	16·5	18·7

53. For Group I the average range was 12·5, or 20·8% of the maximum; for Group II the average range was 16·5, or 27·5% of the maximum; and taking both groups together we find an average range of 18·7, or 31·2% of the maximum. It is true that there is one candidate (No. 8) for whom the extreme difference between the thirteen examiners is only 9 marks on the maximum of 60; but for Candidate No. 10 the extreme range goes up to 28, or over 46% of the maximum. One may well question the validity of a system of test which yields results so discrepant. Qn. 8 is one on which high average marks are earned, and even when they are divided by 4, the differences may affect the fate of a candidate seriously.

54. Qn. 12 is one on a set book in Latin verse, also with a maximum of 60. The marks are set out in Table 19 below :—

TABLE 19

(*Directions the same for both Groups*)

LATIN II. *Qn.* 12 (*Translation of Virgil—about 22 lines*)

Maximum = 60 *marks*

Cand.	*Group I* Examiner						Range	*Group II* Examiner							Range	Extreme Range
	A	B	C	D	E	F		G	H	J	K	L	M	N		
1	27	32	45	27	32	30	18	33	40	41	35	44	35	27	17	18
2	32	40	47	26	35	41	21	29	40	48	36	45	34	33	19	22
3	34	40	44	36	38	39	10	28	42	55	30	52	43	42	27	27
4	8	27	26	16	24	19	19	18	21	35	18	20	21	23	17	27
5	23	39	42	26	35	36	19	40	35	50	31	48	32	42	19	27
6	32	40	40	29	36	36	11	35	37	47	35	49	38	41	14	20
7	40	45	51	42	46	44	11	46	43	54	47	52	47	48	11	14
8	55	56	59	51	53	49	10	45	51	57	52	54	52	53	12	14
9	10	33	35	22	31	29	25	32	31	47	24	41	30	25	23	37
10	35	40	43	45	41	39	10	29	38	50	32	42	34	38	21	21
11	15	26	33	15	20	14	19	29	22	44	24	36	19	23	25	30
12	43	40	53	46	40	44	13	33	47	55	41	47	41	38	22	22
13	43	37	42	36	28	40	15	38	40	50	35	49	35	38	15	22
14	7	27	31	13	19	22	24	23	26	43	10	35	23	23	33	36
15	8	20	20	8	13	12	12	16	18	37	13	31	15	15	24	29
Averages	27·5	36·1	40·7	29·2	32·7	32·9	15·8	31·6	35·4	47·5	30·9	43·0	33·3	33·9	19·9	24·4

55. The results are of the same character as those for Qn. 8, with the discrepancies somewhat increased. For Group I the average range is 15·8, or 26·3% of the maximum ; for Group II the average range is 19·9, or 33·2% of the maximum ; and for the whole thirteen examiners the average range is 24·4, or 40·7% of the maximum. The lowest extreme range is 14 (for Candidates Nos. 7 and 8). There are seven candidates for whom the extreme range is 27 or over ; and there are two candidates (Nos. 9 and 14) for whom the extreme ranges are 36 and 37, i.e. more than 60% of the maximum.

56. We now turn to Qn. 7 (two unseen passages, one in Latin prose, one in Latin verse, for translation into English) in regard to which a difference of opinion of the examiners led to the adoption of two marking-schemes, a more detailed one for Group I (six examiners), and a less detailed one for Group II (seven examiners).[1] Group I had nine more detailed instructions (out of eleven) than Group II. The marks are set out in Table 20 below. The question was attempted by only fourteen candidates out of the fifteen.

[1] More detailed instructions for Group I were also given in respect of the passage of continuous English prose numbered as Qn. 6. But as this was attempted by only one candidate we have left it out of consideration. The range for Group I was 13 (out of 50) and for Group II, 19 out of 50.

TABLE 20

LATIN I. *Qn.* 7 (*Translation of two passages* (*one prose, one verse*) *from Latin into English*)
Marks of 14 *Candidates. Maximum for the question* = 60. (*Group I had the more detailed instructions*)

			GROUP I							GROUP II								
Range of Group I	Examiner	A	B	C	D	E	F	G	H	J	K	L	M	N	Range of Group II	Range of Group II omitting J	Extreme Range for both Groups
4	Cand. 1	6	10	10	6	10	7	11	3	17	6	12	6	7	14	9	14
8	2	7	15	8	15	14½	12	13	6	28	14	18	12	10	22	12	22
12	3	2	14	12	5	11½	9	9	3	25	4	18	12	12	22	15	23
12	4	10	18	6	10	12	10	14	3	28	4	19	15	13	25	16	25
18	5	9	22	21	4	14	20	17	13	30	19	26	17	17	17	13	26
8	6	9	15	7	7	13½	13	22	4	21	9	22	13	18	18	18	18
8	7	1	9	6	1	5	5	10	5	8	5	13	8	11	8	8	12
10	9	6	11	16	7	9	11	12	5	24	6	14	13	18	19	13	19
8	10	7	13	12	5	8	11	13	3	23	3	10	11	22	20	19	20
7	11	11	15	8	10	8½	14	28	7	28	11	15	15	16	21	21	21
6	12	1	7	3	2	3	4	13	0	12	1	3	3	6	13	13	13
10	13	8	15	7	5	8	11	9	0	19	3	10	7	14	19	14	19
9	14	2	11	8	3	6½	6	17	5	23	1	24	9	11	23	23	23
12	15	5	15	10	3	11	14	15	2	22	3	16	14	21	20	19	20
Average 9·4		6·0	13·6	9·6	5·9	9·6	10·5	14·5	4·2	22·0	6·4	15·7	11·1	14·0	18·6	15·2	19·6

57. The average range for Group I is 9·4, or 15·7% of the maximum; for Group II, 18·6, or 31·1% of the maximum. If J's marks are omitted, the average is 15·2, or 25·3% of the maximum. It might seem reasonable, at first sight, to attribute the greater concurrence of the marks for Group I to their more detailed instructions. But as against this hypothesis the following point may be urged. In Table 15, which shows *inter alia* the average ranges expressed as percentages of the maxima, these figures for Group I are, with four exceptions, lower than for Group II. On the whole, we should infer then that there was greater concurrence in the marks of Group I examiners than in those of the other Group. The difference between the average range percentages in the case of Qn. 7 is 15·4, but there is a difference in the same sense of 14·7 (nearly as great) in the case of Qn. 13, and there are large differences in the case of Qns. 8 and 12 (6·6 and 6·9 respectively). Since quite large differences are observed in the case of questions where the instructions to examiners were identical, it would be unjustifiable to attribute the large difference observed in the case of Qn. 7 entirely to the fact that in this case the instructions to examiners were different, though it may be that part of this difference is due to this cause.

CHAPTER III

MARKING OF SCHOOL CERTIFICATE FRENCH SCRIPTS

58. *Character of the Examination Papers.*—The scripts investigated were written in answer to two two-hour papers.

Paper I comprised :—(1) A piece of dictation ;

(2) (a) and (b) Two pieces of French prose for translation, each of approximately 20 lines, and (c) a short piece of French verse for translation.

Paper II comprised :—(1) A piece of English for translation into French ;

(2) An outline in English of a story to be expanded in French, as an exercise in French composition.

59. *Special Object of the Investigation.*—The object of the investigation was to inquire into the degree of consistency in the marking of a set of scripts by the members of two independent Boards of Examiners, all accustomed to the same standards and to team-work on well-established lines.

60. *Selection of Examiners.*—Two Boards were set up, each consisting of a Chief Examiner and six other examiners, selected, with the concurrence of the Chief Examiner, from the panel of a single School Certificate authority, other than that by whom the scripts were furnished. The two Chief Examiners belonged to the same panel as all the other examiners.

61. *Selection of Scripts.*—A School Certificate authority furnished us with a complete mark-list of the candidates at an examination. On the basis of this list we asked for 150 scripts of candidates, selected so as to provide " trial scripts " and " scripts for final marking " in the manner set out below.

62. *Trial Scripts.*—For the purpose of serving as trial scripts, thirty-three scripts were selected of which the original total

marks varied from 77 to 87; thirty-four scripts were selected of which the original total marks varied from 49 to 50; and thirty-three scripts were selected of which the original total marks varied from 10 to 16. These scripts were so chosen with a view to supplying to each examiner as trial scripts two "good," two "medium," and two "poor" specimens. Samples chosen as trial scripts were supplied in original to examiners, according to the procedure described. All indications of the origin of the scripts and of any previous corrections were entirely removed from all the scripts.

63. *Scripts for Final Marking.*—For final marking, fifty scripts were selected in such a way that the marks allotted to them by the original authority corresponded roughly to a "normal" frequency distribution. The total range of their marks was from 2 to 87. These scripts were arranged in random order, re-numbered, and, after the removal of all indications of their origin and previous marking, were reproduced photographically. The photographic reproductions were excellent and were as easy to mark as the originals. Identical copies were supplied to the examiners for final marking after the preliminary process of the Board concerned had been carried out as described below.

64. *Marking-schemes.*—Each Chief Examiner was furnished with a number of trial scripts, and, on the basis of the examination-papers and these scripts, drew up a detailed marking-scheme which was discussed and settled at a meeting of his Board, held at the offices of the Examinations Enquiry Committee. A number of scripts were on the table at the meeting, which in each case lasted several hours. The main outlines of the two marking-schemes are set out below.

				Board I	*Board II*
				Maximum Marks	Maximum Marks
Paper I	Question	1	Dictation	11	20
		2a	Translation	15	35
		2b	,,	20	30
		2c	,,	15	15
			Style	—	10
Paper II	Question	1		30	55
		2		20	35
			Total	111	200

The above marks were subsequently reduced in both cases to a maximum of 100.

In addition to the schemes summarised above, Board I adopted five " General Rules " applicable to both Papers, and schemes of detailed directions including about 640 items for Paper I and 290 for Paper II.

Board II adopted a set of General Rules applicable to the questions in Paper I and another set for Paper II, and in addition to these General Rules schemes of detailed directions including 700 items for Paper I and 300 for Paper II. These detailed directions did not, of course, require any appreciable effort of memory on the part of the examiners ; the majority concerned the rendering into French or English, as the case might be, of particular words or phrases in the passages set for translation.

65. It is interesting to note that while the general methods used by the two Boards were obviously the same, the detailed directions were in a number of cases different, and in some actually conflicting.

For Paper I there were 35 conflicting directions, and for Paper II there were 16, i.e. 51 conflicting directions in all, out of a total possible of between 900 and 1,000. The differences mainly concerned points of English and French style in translation. It would be difficult to maintain, in view of the large number of cases, that any single candidate would be likely to suffer severely from unmerited blows of chance on account of these differences. He might be as likely to gain in one case as to lose in another—though Board II was on the whole more severe. It disallowed 36 renderings permitted by Board I, whereas Board I only disallowed 15 renderings permitted by Board II. In a very large number of the passages discussed there were a number of alternatives allowed by both Boards.

66. After the examiners' meeting, each member of the Board received six trial scripts (two good, two medium, two bad ; see para. 62 above) and marked them for the scrutiny of the Chief Examiner, by whom they were returned with his comments. The Chief Examiner of one Board amended his marking-scheme somewhat after receiving the marked trial scripts.

67. Copies of the fifty scripts reproduced photographically were then circulated to each examiner and a week was allowed for marking. After the examiners had marked these scripts, eight or nine of the marked scripts were chosen from the fifty corrected by each examiner in such a way as to secure the presence of one script of each candidate in the new set of fifty marked scripts, and the whole of this set so chosen was sent

to the Chief Examiner. The Chief Examiner, on the basis of the marks of these scripts, then prescribed certain adjustments in the marks of individual members of the Board, which were made in the office, and of which details are given below.

Board I.—The marks of Examiners A and E were reduced by 1%, those of C and F by 2%, of D by 3%, of B by 4%.

Board II.—The marks of Examiner J were reduced by 2 units throughout, and those of L were raised by 2 units. The marks of H were reduced in accordance with the following scale: the marks from 50 to 100 were reduced by 8; those from 45 to 49 by 7; from 40 to 44 by 6; from 35 to 39 by 5; from 10 to 34 by 4; below 10 no change. The marks of Examiners G, K and M were not altered.

68. After the procedure adopted in the foregoing paragraph had been carried out, a list of the adjusted marks of the candidates, 300 marks in all, was drawn up for each Board, arranged in each case in order of magnitude; and the list was sent to the relevant Chief Examiner. The examiners had been asked at the outset to bear in mind, while marking, the awards of Failure, Pass, Credit and Distinction, based on the following limiting marks: Pass 35%, Credit 45%, Distinction 70%.

After a scrutiny of their respective lists, the two Chief Examiners, with the object of securing what seemed to them a reasonable number of awards in the various categories, finally fixed their limits as follows:—

Board I ..	Pass 33%;	Credit 43%;	Distinction 67%
Board II ..	Pass 35%;	Credit 50%;	Distinction 70%

69. The adjusted marks are set out in Table 21 below, and symbols showing the awards by the Examiners of Failure marks (F), Pass (P), Credit (C) and Distinction (D) are set out in Table 21A below.

70. It will be noticed that the range of marks for individual candidates (i.e., the difference between the highest and lowest marks allotted to them) varies with Board I from 2 to 19; and with Board II from 3 to 16.

The average range for Board I is 10·6, for Board II 7·8. But it is to be pointed out that, in 32 cases out of 50, Examiner A marks lower than any other member of his Board; if his results were for the moment excluded, the range of his Board would only be from 2 to 17, and the average range would be 7·5, approximately the same as, but slightly lower than, that of Board II.

TABLE 21

Total Adjusted Marks

Cand.	*Board I* Examiner A	B	C	D	E	F	Range	Range irrespective of Examiner A's marks	*Board II* Examiner G	H	J	K	L	M	Range
1	65	66	65	62	57	68	11	11	65	53	68	66	64	69	16
2	52	54	51	49	50	48	6	6	49	45	53	55	50	52	10
3	46	50	50	47	44	48	6	6	46	47	54	55	50	48	9
4	35	44	46	41	40	40	11	6	38	42	43	42	39	42	5
5	13	23	25	29	18	20	16	11	13	21	20	16	19	18	8
6	44	57	53	54	55	51	13	6	52	59	58	59	58	64	12
7	58	60	62	59	60	66	8	7	63	60	64	59	70	69	11
8	26	36	37	38	31	30	12	8	37	46	36	40	44	38	10
9	6	16	15	17	18	12	12	6	7	7	10	7	11	8	4
10	46	61	59	56	53	63	17	10	60	59	58	58	61	58	3
11	47	52	50	48	47	53	6	6	49	55	56	50	49	51	7
12	66	70	61	61	60	67	10	10	61	62	62	66	60	59	7
13	42	51	51	52	48	54	12	6	55	57	50	51	54	49	8
14	28	36	32	32	34	31	8	5	29	34	25	26	29	24	10
15	35	49	42	43	46	47	14	7	46	52	50	47	51	46	6
16	31	39	32	37	36	32	8	7	33	38	31	36	31	27	11
17	1	7	6	5	8	5	7	3	4	4	0	2	3	2	4
18	66	64	62	63	64	73	11	11	63	67	70	66	68	67	7
19	68	70	68	72	70	68	4	4	67	69	73	70	67	72	6
20	30	43	35	42	39	43	13	8	41	43	42	41	42	38	5
21	63	63	68	63	64	65	5	5	60	65	59	67	61	65	8
22	49	57	53	59	46	53	13	13	48	52	56	53	49	50	8
23	48	56	50	49	55	51	8	7	50	56	53	55	58	54	8
24	50	54	54	54	54	61	11	7	55	56	50	58	56	52	8
25	67	68	68	66	67	68	2	2	68	63	70	68	70	71	8
26	30	38	38	41	40	40	11	3	37	41	39	38	39	35	6
27	28	36	30	33	37	37	9	7	32	39	37	39	35	33	7
28	32	42	40	43	36	40	11	7	40	42	43	48	45	42	8
29	31	41	39	41	38	39	10	3	41	41	36	44	40	39	8
30	42	52	49	53	49	48	11	5	51	47	47	54	57	53	10
31	45	52	44	48	50	53	9	9	54	54	57	50	58	55	8
32	36	42	42	46	43	40	10	6	42	44	43	45	42	44	3
33	50	58	47	51	51	48	11	11	50	55	50	51	54	51	5
34	44	53	49	50	50	50	9	4	48	48	49	51	53	47	6
35	39	54	51	48	47	46	15	8	46	48	51	52	52	47	6
36	76	80	72	72	68	75	12	12	68	73	71	77	72	74	9
37	12	23	18	21	19	21	11	5	20	28	21	18	24	15	13
38	20	39	28	34	31	32	19	11	26	31	26	29	32	26	6
39	35	48	45	41	39	38	13	10	34	38	30	43	36	37	13
40	17	26	24	23	21	23	9	5	14	16	14	18	18	12	6
41	66	74	65	66	66	71	9	9	66	63	65	70	65	62	8
42	44	54	48	54	50	51	10	6	51	45	50	55	53	47	10
43	45	62	45	50	55	52	17	17	58	50	52	50	57	52	8
44	49	56	56	51	59	65	16	14	56	59	60	56	60	61	5
45	41	51	55	55	50	60	19	10	59	58	65	57	58	56	9
46	70	72	70	71	73	77	7	7	68	65	73	74	69	71	9
47	48	55	53	55	55	58	10	5	58	60	54	54	60	57	6
48	84	79	83	83	81	83	5	4	80	77	80	83	80	78	6
49	16	24	26	24	21	26	10	5	21	24	22	21	24	18	6
50	55	61	54	49	54	64	15	15	55	54	56	52	63	52	11
Averages	42·7	50·4	47·3	48·0	46·9	49·1	10·6	7·5	46·7	48·2	48·0	48·8	49·2	47·1	7·8

TABLE 21A

AWARDS OF INDIVIDUAL EXAMINERS CALCULATED FROM ADJUSTED MARKS

Cand.	*Board I* Examiner						*Board II* Examiner					
	A	B	C	D	E	F	G	H	J	K	L	M
1	C	C	C	C	C	D	C	C	C	C	C	C
2	C	C	C	C	C	C	P	P	C	C	C	C
3	C	C	C	C	C	C	P	P	C	C	C	P
4	P	C	C	P	P	P	P	P	P	P	P	P
5	F	F	F	F	F	F	F	F	F	F	F	F
6	C	C	C	C	C	C	C	C	C	C	C	C
7	C	C	C	C	C	C	C	C	C	C	D	C
8	F	P	P	P	F	F	P	P	P	P	P	P
9	F	F	F	F	F	F	F	F	F	F	F	F
10	C	C	C	C	C	C	C	C	C	C	C	C
11	C	C	C	C	C	C	P	C	C	C	P	C
12	C	D	C	C	C	D	C	C	C	C	C	C
13	P	C	C	C	C	C	C	C	C	C	C	P
14	F	P	F	F	P	F	F	F	F	F	F	F
15	P	C	P	C	C	C	P	C	C	P	C	P
16	F	P	F	P	P	F	F	P	F	P	F	F
17	F	F	F	F	F	F	F	F	F	F	F	F
18	C	C	C	C	C	D	C	C	D	C	C	C
19	D	D	D	D	D	D	C	C	D	D	C	D
20	F	C	P	P	P	C	P	P	P	P	P	P
21	C	C	D	C	C	C	C	C	C	C	C	C
22	C	C	C	C	C	C	P	C	C	C	P	C
23	C	C	C	C	C	C	C	C	C	C	C	C
24	C	C	C	C	C	C	C	C	C	C	C	C
25	D	D	D	C	D	D	C	C	D	C	D	D
26	F	P	P	P	P	P	P	P	P	P	P	P
27	F	P	F	P	P	P	F	P	P	P	P	F
28	F	P	P	C	P	P	P	P	P	P	P	P
29	F	P	P	P	P	P	P	P	P	P	P	P
30	P	C	C	C	C	C	C	P	P	C	C	C
31	C	C	C	C	C	C	C	C	C	C	C	C
32	P	P	P	C	C	P	P	P	P	P	P	P
33	C	C	C	C	C	C	C	C	C	C	C	C
34	C	C	C	C	C	C	P	P	P	C	C	P
35	P	C	C	C	C	C	P	P	C	C	C	P
36	D	D	D	D	D	D	C	D	D	D	D	D
37	F	F	F	F	F	F	F	F	F	F	F	F
38	F	P	F	P	F	F	F	F	F	F	F	F
39	P	C	C	P	P	P	F	P	F	P	P	P
40	F	F	F	F	F	F	F	F	F	F	F	F
41	C	D	C	C	C	D	C	C	C	D	C	C
42	C	C	C	C	C	C	C	P	C	C	C	P
43	C	C	C	C	C	C	C	C	C	C	C	C
44	C	C	C	C	C	C	C	C	C	C	C	C
45	P	C	C	C	C	C	C	C	C	C	C	C
46	D	D	D	D	D	D	C	C	D	D	C	D
47	C	C	C	C	C	C	C	C	C	C	C	C
48	D	D	D	D	D	D	D	D	D	D	D	D
49	F	F	F	F	F	F	F	F	F	F	F	F
50	C	C	C	C	C	C	C	C	C	C	C	C

71. More important to a candidate than his numerical marks is the question whether he is reported as having been classed under one of the headings, Failure, Pass, Credit, or Distinction ; we now deal with this matter. The distributions of the awards made by the several examiners of the two Boards were as follows :—

TABLE 22

	Board I Examiner						*Board II* Examiner					
	A	B	C	D	E	F	G	H	J	K	L	M
Failures	15	6	10	7	8	10	11	8	10	8	9	10
Passes	8	9	7	9	9	7	14	16	10	11	11	14
Credits	22	28	27	30	28	24	24	24	24	26	26	21
Distinctions	5	7	6	4	5	9	1	2	6	5	4	5
Total	50	50	50	50	50	50	50	50	50	50	50	50

Thus Examiner A gave a Failure mark to fifteen candidates, a Pass to eight, Credit to twenty-two, and Distinction to five candidates.

Table 22 gives an indication of the differences of standard of the different examiners, a point which will be dealt with more in detail later. It will be noticed that the number of Failures varies from 6 to 15 ; of Passes from 7 to 16 ; of Credits from 21 to 30 ; and of Distinctions from 1 to 9 : very remarkable differences.

72. But one of the questions of most importance is that of the agreement or difference in regard to the awards to individual candidates, a point on which statistics of this kind give no information (see para. 10 relating to the investigation on School Certificate History Scripts).

Among the members of the same Board there is complete agreement in regard to rather over half the candidates in each case. The following Table shows the extent of agreement among the members of each of the two Boards.

Number of Cases in which the Examiners make the same Award

	Board I	*Board II*
6 examiners make the same award	27	30
5 examiners make the same award, 1 differs	10	5
4 examiners make the same award, 2 agree on another award	9	9
4 examiners make the same award, 2 differ	1	–
3 examiners make the same award, 3 agree on another award	2	6
3 examiners make the same award, 2 agree on another award, and 1 differs	1	–
	50	50

Thus the six examiners of Board I agree on the award to twenty-seven candidates, and those of Board II agree as to thirty candidates. Candidate No. 20, as to whose fate there is most disagreement on the part of the examiners of Board I, is given the following marks by the examiners of that Board: 30, 43, 35, 42, 39, 43, the awards being F, C, P, P, P, C. This candidate is given 41, 43, 42, 41, 42 and 38 marks by the examiners of Board II, representing in each case a Pass.

It seems surprising that, after the various processes of agreeing on marking-schemes, and checking by the Chief Examiners of the individual examiners' methods of marking, agreement is only reached as to the fate of just over half the candidates.

73. The candidates about whom there is agreement between the members of a Board are distributed throughout the whole range :—

Board I	Candidates	*Board II*	Candidates
Agree in "ploughing"	6	Agree in "ploughing"	8
,, ,, giving Credit to	17	,, ,, passing	7
,, ,, giving Distinction to	4	,, ,, giving Credit to	14
		,, ,, giving Distinction to	1
	27		30

74. But it will be seen that the number of Failure marks given by the members of Board I varies from 6 to 15, with an average of 9; those given by the members of Board II vary from 8 to 11, with the same average of 9. The number of Passes for Board I varies from 7 to 9, with an average of 8; for Board II the corresponding figures vary from 10 to 16, with an average of 13. The number of Credits for Board I varies from 22 to 30, with an average of 27; for Board II the figures vary from 21 to 26, with an average of 24. The number of Distinctions for Board I varies from 4 to 9, with an average of 6; for Board II the figures vary from 1 to 6, with an average of 4.

75. To the four candidates to whom the examiners of Board I are unanimous in awarding Distinction, the following awards are made by the examiners of Board II :—

One candidate is given Distinction by six examiners.

One candidate is given Distinction by five examiners and Credit by one examiner.

Two candidates are given Distinction by three examiners and Credit by three examiners.

The awards of the examiners of Board II are on the whole lower than those of Board I.

To the seventeen candidates to whom the examiners of Board I are unanimous in awarding a Credit, the following awards are made by the examiners of Board II :—

One candidate is given Credit by five examiners and Distinction by one examiner.

Ten candidates are given Credit by six examiners.

Four candidates are given Credit by four examiners and Pass by two examiners.

One candidate is given Credit by three examiners and Pass by three examiners.

One candidate is given Credit by two examiners and Pass by four examiners.

Thus, on the whole, the examiners of Board II gave these candidates a lower award than that given by the examiners of Board I.

Further, to the fourteen candidates to whom the examiners of Board II are unanimous in awarding a Credit, the following awards are made by Board I :—

One candidate is given a Pass by one examiner and Credit by five examiners.

Ten candidates are given Credit by six examiners.

Two candidates are given Credit by five examiners and Distinction by one examiner.

One candidate is given Credit by four examiners and Distinction by two examiners.

Here, for the Credits awarded by Board II, the members of Board I more frequently substitute Distinctions than Passes.

Any large difference between the distributions of awards would necessarily be unlikely with two Chief Examiners acting in accordance with the same traditions.

76. We turn now to the question of numerical marks. A glance at Table 21 (page 38) shows how great may be the discrepancies between individual examiners which survive the minutely detailed marking-schemes set up by both Boards. We summarise below the extreme differences (or " ranges ") of marks allotted to individual candidates by the examiners on the two Boards.

TABLE 23

DISTRIBUTION OF THE RANGE OF MARKS

Range	*Board I*	*Board II*
2	1	–
3	–	2
4	1	2
5	2	4
6	3	10
7	2	4
8	4	12
9	5	4
10	6	5
11	9	3
12	4	1
13	4	2
14	1	–
15	2	–
16	2	1
17	2	–
18	–	–
19	2	–
Total	50	50
Average Range	10·6	7·8

77. The case of Examiner A is particularly interesting. After seeing nine of his corrected scripts the Chief Examiner came to the conclusion that they were too high by 1% (see para. 67 above) and in Table 21 above of the adjusted marks they have been reduced accordingly. But they prove to be lower than those of any of his colleagues in 32 cases out of 50. The difference of 1% is in itself hardly significant.

78. The examiners of Board II mark closer together than do those of Board I. In the case of Board I there are thirteen candidates with a range of 13 marks or more, that is, more than a quarter of the candidates ; and there are twenty-four candidates, nearly half, with a range of 9 to 12 inclusive. Remembering that the difference between Pass and Credit is the difference between 33 and 43 marks for this Board, we can appreciate what the average range of marks (10·6) means to the candidates examined by the members of this Board. The average range is the span between Pass and Credit. For Board II the average range is half the span between Pass and Credit.

In spite of all the care taken to eliminate the personal equation of the examiner, this apparently still remains to a considerable extent.

79. We recall the fact that the Chief Examiners adjusted the marks on the basis of the examination of about one-sixth of

the total number of scripts (eight or nine) in each case. The following Table shows how the averages were affected by the adjustment.

Board I				*Board II*			
Examiner	Average of Total Marks	Adjustment made by Chief Examiner	Finally[1] adjusted Average	Examiner	Average of Total Marks	Adjustment made by Chief Examiner	Finally[1] adjusted Average
A	42·9	−0·6	42·3	G	46·4	0	46·4
B	52·4	−2·1	50·3	H	55·2	−7·2	48·0
C	48·2	−1·0	47·2	J	49·6	−2·0	47·6
D	49·3	−1·5	47·8	K	48·5	0	48·5
E	47·3	−0·5	46·8	L	47·1	+2·0	49·1
F	49·9	−1·0	48·9	M	46·9	0	46·9
Average	48·3		47·2		48·9		47·7

80. It is obvious that the sample on which the Chief Examiner based his adjustment of A's marks was not typical. Instead of his marks being lowered they should have been raised so as to yield an average of about 48%. In all the other cases the adjustment has been quite reasonably satisfactory. No Chief Examiner and no authority would expect the average of each examiner for fifty scripts to be *exactly* the same. The adjustment has broken down in one case out of twelve.

81. It should be pointed out here that no mere adjustment of averages will of itself remove discrepancies between the distribution of awards by individual examiners (see para. 71 above) since the differences in marks depend not only on differences of standard from examiner to examiner, but on "random" departures from those standards and on differences of spreading.

82. Turning now to the marking of the answers to the six separate questions, we subjoin in Tables 24 and 25 below an analysis of the marks allotted to the fifty candidates by the twelve examiners :—

TABLE 24
Board I

(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			Paper I			Paper II			
		Dictn.	2a	2b	2c	1	2	Total	Per cent.
(*a*) Maximum Marks		11	15	20	15	30	20	111	100
	Examiner								
	A	8·80	6·86	9·76	8·52	7·06	6·64	47·64	42·9
	B	8·91	7·76	10·42	10·28	9·66	11·16	58·19	52·4
(*b*) Average Marks	C	9·18	7·06	9·92	10·32	8·58	8·50	53·56	48·3
	D	8·71	7·64	10·70	10·16	7·60	9·94	54·75	49·3
	E	9·13	6·64	9·56	9·72	8·40	9·06	52·51	47·3
	F	9·11	7·62	10·84	10·46	8·90	8·48	55·41	49·9

[1] These figures differ slightly from those in Table 21, as they are obtained by subtraction of the average adjustments from the averages of the original marks reduced to a percentage form, whereas those in Table 21 are the averages of the adjusted marks.

(1)		Paper I (2) Dictn.	(3) 2a	(4) 2b	(5) 2c	Paper II (6) 1	(7) 2
(c)	Average of Examrs.' Averages	8·97	7·26	10·20	9·91	8·37	8·96
(d)	% of Maximum mark	81·5	48·4	51·0	66·1	27·9	44·8
(e)	Highest Av. Marks	9·18	7·76	10·84	10·46	9·66	11·16
	Lowest Av. Marks	8·71	6·64	9·56	8·52	7·06	6·64
(f)	Difference	0·47	1·12	1·28	1·94	2·60	4·52
(g)	% of Maximum Mark	4·3	7·5	6·4	12·9	8·7	22·6
(h)	Average Range	1·02	2·54	3·88	3·20	3·80	5·64
(j)	% of Maximum Mark	9·3	16·9	19·4	21·3	12·7	28·2

TABLE 25

Board II

(1)		(2) Dictn.	Paper I (3) 2a	(4) 2b	(5) 2c	(5a) Stylo	Paper II (6) 1	(7) 2	(8) Total	(9) Per cent.
(a)	Maximum Marks	20	35	30	15	10	55	35	200	100
	Examiner									
	G	13·54	17·26	7·42	5·68	5·76	27·20	15·96	92·82	46·4
	H	13·58	19·38	13·96	8·70	6·00	29·50	19·30	110·42	55·2
(b) Average	J	13·88	18·06	10·26	6·98	4·40	29·06	16·56	99·20	49·6
Marks	K	13·30	19·20	9·10	7·30	5·58	28·14	14·48	97·10	48·5
	L	13·42	17·62	9·06	6·46	4·86	29·52	13·36	94·30	47·1
	M	13·24	16·66	9·24	5·18	4·44	27·58	17·40	93·74	46·9
(c)	Average of Exmrs.' Averages	13·49	18·03	9·84	6·72	5·17	28·50	16·18		
(d)	% of Maximum Mark	67·5	51·5	32·8	44·8	51·7	51·8	46·2		
(e)	Highest Av. Marks	13·88	19·38	13·96	8·70	6·00	29·52	19·30		
	Lowest Av. Marks	13·24	16·66	7·42	5·18	4·40	27·20	13·36		
(f)	Difference	0·64	2·72	6·54	3·52	1·60	2·32	5·94		
(g)	% of Maximum Mark	3·2	7·8	21·8	23·5	16·0	4·2	17·0		
(h)	Average Range	1·24	5·04	7·30	4·54	2·78	4·52	9·04		
(j)	% of Maximum Mark	6·2	14·4	24·3	30·3	27·8	8·2	25·8		

83. The above Tables give the average marks, which serve to indicate standards of marking, and the average range of marks, these differences of marking of the same script being due to the examiners having different standards and to the inevitable variations from the standard, described as random marking.

As a general remark we may point out that where the averages of the examiners of a Board are close this shows the control of the Chief Examiner's marking-scheme; where they differ the individual examiners have escaped, no doubt involuntarily, from this control, or the marking-scheme has itself been defective.

84. Before entering into details we must point out that the maxima allotted by the two Boards to different parts of the question-papers differ, but are not very different. They are shown in Table 26 below.

TABLE 26

MAXIMUM MARKS PER QUESTION AS PERCENTAGES OF TOTALS

	Paper I					Paper II	
Question	1 (Dictation)	2a	2b	2c	Style (for 2a, 2b and 2c)	1	2
Board I	9·9	13·5	18·0	13·5		27·0	18·0
			45				
Board II	10·0	17·5	15·0	7·5	5·0	27·5	17·5
			45				

The marks for style, given by Board II only, are allotted in respect of answers to the *three* questions 2a, 2b, and 2c, so that no exact comparison between the separate maxima assigned by the two Boards for these three questions is possible.

The chief difference relates to the maximum allotted to Qn. 2c. But any difference between the Boards in respect of the relative importance of the questions is entirely eclipsed by their differences in regard to the relative importance of the answers.

85. It is interesting, neglecting individual differences, to compare for each question the averages of the average marks of the examiners of Board I with the corresponding averages of the examiners of Board II, these averages being expressed as percentages of the maxima obtainable in each case (see Tables 24 and 25, row *d*, in each case).

TABLE 27

AVERAGE MARKS EXPRESSED AS PERCENTAGES OF THE MAXIMUM IN EACH CASE[1]

Question	Paper I					Paper II	
	1 (Dictation)	2a	2b	2c	Style	1	2
Board I	81·5	48·4	51·0	66·1		27·9	44·8
Board II	67·5	51·5	32·8	44·8	51·7	51·8	46·2
Difference	14·0	3·1	18·2	21·3		23·9	1·4

86. Table 27 above is probably the most important feature in the whole analysis. In the case of only two questions were the standards approximately the same, Qn. 2a of Paper I (a piece of translation from French into English) and Qn. 2 of Paper II (original composition in French). The averages for the other questions differ widely. It is strange that in Qns. 2b and 2c, both pieces of translation from French into English, the differences should be so much greater than for the other piece of translation from French into English, Qn. 2a. The elaborate systems adopted yield not only different results with examiners of the same Board, but widely different average results for the two Boards.

87. The difference between the two Boards in dealing with Dictation (for which Board I gave an average of 81·5% as against the 67·5% of Board II) is almost as large as in dealing with translation from French into English ; while the largest difference of all is to be found in the marking of the piece of translation from English into French (Paper II, Qn. 1). Here the average mark of Board I is only 27·9% as against the average of 51·8% awarded by Board II. In regard to these two points of Dictation and translation from English into French, the standards of the two Boards may be fairly described as irreconcilable. Board II gives nearly twice as many marks on the average as Board I to the same translations. It is quite obvious that candidates weak in translation from English into French have been dealt with much more severely by Board I than by Board II ; while candidates weak in Dictation have been dealt with more severely by Board II than by Board I.

[1] It is to be noted as a detail which does not seriously affect the general conclusions to be drawn from the Table that whereas Board II allotted 10 marks (out of 110) for style (including spelling, tidiness, English style, etc.) in the passages for translation from French into English in Qns. 2a, 2b, and 2c, Board I used a different system, so that the marks for 2a, 2b, 2c are not completely comparable. The addition of 10 % to the marks of Board II for Qns. 2a, 2b, 2c would bring these averages up to 56·6, 36·1 and 49·3 % respectively.

88. If we tried to ascertain by experiment what the examining authorities abstain from telling us, viz., what a candidate who just passes their examination *can do*, we should find a very different answer if we investigated the capacities of those who just pass with Board I from what we should obtain if we investigated the capacities of those who just pass with Board II.

89. An answer to Qn. 1 of Paper II is a more important factor in deciding the result for Board II than for Board I; on the other hand for Qn. I, 2b the reverse is the case. The awards indicated by the average marks of the six examiners of each Board, showing performance on these two questions, are given below, together with the average awards of the two Boards on the whole examination (see para. 74)[1] :—

Board I	D	C	P	F
Question I, 2b	8	24	9	9
Question II, 1	1	9	9	31
Average Distribution	6	27	8	9
Board II	D	C	P	F
Question I, 2b	1	16	7	26
Question II, 1	5	33	2	10
Average Distribution	4	24	13	9

The distribution of awards showing performance in Qn. II, 1, is nearer to the average distribution in the case of Board II than in the case of Board I, and similarly the distribution of awards showing performance in Qn. I, 2b is nearer to the average distribution in the case of Board I than in the case of Board II.

The low marks awarded to Qn. II, 1 by the examiners of Board I result in more than half the candidates receiving a Failure mark in that question, whereas actually on the whole paper more than half the candidates receive a Credit mark. Similarly, the low marks awarded to Qn. I, 2b by the examiners of Board II result in more than half the candidates receiving a Failure mark in that question, whereas actually on the whole paper almost half the candidates receive a Credit mark.

Putting the matter in another way, we may say that thirty-eight out of the fifty candidates receive a Credit or Distinction mark from the examiners of Board II for their performance on Qn. II, 1, which will help them considerably in their total marks, whereas only ten receive such high marks from the examiners of Board I for answers to this question. Similarly, thirty-two out of the fifty candidates receive Credit or Distinction marks from the examiners of Board I for their performance on Qn. I, 2b,

[1] In this table, the awards, D, C, P, F, are allotted to the individual question, for the same percentages of the maxima respectively as Distinction, Credit, Pass and Failure are allotted to the papers as a whole.

whereas only seventeen receive such high marks from the examiners of Board II for answers to this question.

90. When we consider the average marks for the two Boards of the scripts treated as a whole, these differences are eliminated. But, as we have seen above, the fate of individual candidates depends on differences which the similarity of general results effectively conceals.

91. We have compared in para. 79 above the averages of the different examiners for the scripts as a whole. We find that some examiners are fairly consistent in their standards of marking, as they pass from one question to another. Thus we see from Table 24, rows (*b*) and (*c*), that in Board I, A has an average mark not only lower as a whole but lower for each question than the average of the other examiners ; that B has higher averages than the general one except in Dictation, and that F has a higher average than the general one except in Paper II, Qn. 2. The other examiners, C, D, and E have averages for different questions sometimes higher and sometimes lower than the general averages. In Board II we find from Table 25, rows (*b*) and (*c*), that H gives consistently higher marks than the general average, and J higher marks except for " style " ; and that L and M give on the whole lower marks. G and K vary more.

92. We shall now examine more closely the average marks assigned for the different questions and their ranges, apart from the question of individual examiners. The following figures are abstracted from Tables 24 and 25 above.

TABLE 28

(*All figures are given as percentages of the maxima for the various questions*)

Question	Board I		Board II	
	Difference between highest and lowest averages	Average Range	Difference between highest and lowest averages	Average Range
Paper I				
Dictation	4·3	9·3	3·2	6·2
2a, 2b, 2c — Translation French into English	6·7 / 6·4 / 12·9	16·9 / 19·4 / 21·3	7·8 / 21·8 / 23·5	14·4 / 24·3 / 30·3
Style			16·0	27·8
Paper II				
1 — Translation English into French	8·7	12·7	4·2	8·2
2 — French Composition	22·6	28·2	17·0	25·8

Where the figures in this Table are large it means that there is considerable divergence of opinion amongst the examiners as to the merit of an answer ; where they are small there is considerable agreement.

93. We note first of all the general similarity of the figures in the two halves of Table 28.

With both Boards there is far less difference between the examiners in marking Dictation than in marking any other part of the paper. They have only to compare the candidates' efforts word for word with the original scripts. The difference in the valuations of the two Boards as a whole referred to above (paras. 85 and 86) is due to a difference of method employed, and not to uncertainties of estimation. We noted the very considerable differences between the examiners in each Board in marking translations from French into English, and in particular from French poetry (Qn. I, 2c).

In marking translation from English into French the examiners of Board II are closer together than those of Board I. In marking French composition, on both Boards there are great divergencies.

CHAPTER IV

MARKING OF SCHOOL CERTIFICATE CHEMISTRY SCRIPTS

94. *Character of the Examination Paper.*—The scripts investigated were written in answer to a three-hour paper on elementary chemistry, comprising eight questions, of which the candidates were required to answer six.

95. *Special Object of the Investigation.*—The object of the investigation was identical, except for the difference of subject, with the investigation on School Certificate French (see Chapter III above), i.e. to inquire into the degree of consistency in the marking of a set of scripts by the members of two independent Boards of Examiners, all accustomed to work to the same standards and to team-work on well-established lines.

96. *Selection of Examiners.*—Two Boards were set up, each consisting of a Chief Examiner and six other examiners, selected with the concurrence of the Chief Examiner, from the panel of a School Certificate authority other than that by which the scripts were furnished. The Chief Examiners belonged to the same panel as all the other examiners.

97. *Selection of Scripts.*—A School Certificate authority furnished us with a complete mark-list of the candidates at an examination. On the basis of this list we selected two hundred and fifty scripts of which the marks allotted at the original examination varied from a number approaching zero to one approaching the maximum, so that the marks were distributed approximately in accordance with the " normal frequency curve."

98. *Scripts for Final Marking.*—From the two hundred and fifty scripts, thirty were selected of which the original marks also varied from a number approaching zero to one approaching the maximum, again roughly in accordance with the " normal frequency curve," and, after the removal of all indications of the origin and previous marking, these were reproduced photographically. It may be mentioned that the photographic reproductions of

faint pencil drawings of apparatus, etc., were much clearer than the originals. The reproduction left nothing to be desired, and the copies were as easy to mark as the originals. The only reason for not reproducing fifty scripts, as in the case of the French investigation, was one of expense. The average length of a script was thirteen pages, and the cost of photographing them was considerable.

99. *Trial scripts.*—The trial scripts were selected from the remaining two hundred and twenty scripts in the manner described in para. 102 below. As with the scripts chosen for final marking, every trace of the origin and previous marking of each script was removed before it was used.

100. *Marking-schemes.*—Each Chief Examiner was furnished with a copy of the examination-paper, and with twelve trial scripts, on the basis of which he drew up a detailed marking-scheme, which was discussed and settled at a meeting of his Board held at the offices of the Examinations Enquiry Committee. A number of scripts were on the table at the meeting, which in each case lasted for several hours.

Board I allotted 17 marks to one question (No. 6) and 16 to each of the others; and gave 3 "grace marks" for equations throughout the script, so that the maximum was 100. The number of detailed directions was about 85.[1]

Board II allotted 17 marks to each question so that the maximum was 102, and the number of detailed directions was about 95.[1]

Although Board II gave more detailed instructions in the aggregate, Board I gave much more detailed instructions in regard to certain questions, e.g. Qn. 7.

For certain details the two Boards gave the same marks. On one question Board II gave three marks each for two definitions, while Board I gave only 1 mark each, a difference of 4 marks for identical answers.

101. An inspection of the two marking-schemes as a whole showed that there was a common general tradition behind them, but also, in spite of this common tradition, considerable variety of detail in regard to the relative importance of different pieces of information; and the differences between the Boards would probably have been much greater but for the fact that the

[1] An estimate of the number of detailed instructions can only be approximate; it may be difficult to decide whether an instruction referring to two parts of the same answer, or to two or more compounds, should be reckoned as one "detailed direction" or more.

instruction to candidates to select any six questions out of the eight made it necessary to allot identical or almost identical maxima to each question.

102. After the examiners' meeting, each member of the relevant Board received six trial scripts, two good, two bad, and two medium, which he marked at home and returned for the criticism of the Chief Examiner.

Copies of the thirty scripts reproduced photographically were then circulated to each examiner and a week was allowed for marking. The Chief Examiner of Board II suggested (after Board I had finished their work) that he also should mark thirty scripts to give him the necessary experience for the revision of the marked scripts. His request was acceded to, and he was supplied with thirty scripts other than those duplicated and comprising ten good, ten bad, and ten medium.

103. *Adjustment of Marks.*—On the completion of the marking of the thirty duplicated scripts, the Chief Examiner of each Board received nine of these scripts from each member : after scrutiny of these scripts, he decided whether the examiner in question had observed the required standard of marking and what adjustments, if any, should be made in his aggregate marks.

In the case of Board I, these nine scripts were sent direct to the Chief Examiner by the examiners concerned, the marks ranging roughly from 40 to 70. In the case of Board II, the nine scripts were chosen in the office so as to give representative samples to the Chief Examiner. The instructions given by the Chief Examiners for the adjustment of the marks of the different examiners were as follows :—

Board I		*Board II*	
Examiner A	Add 5% to each mark	Examiner G	No change
,, B	Subtract 5% from each mark	,, H	Raise by 5 marks
,, C	Add 5% to each mark	,, J	Lower by 5 marks
,, D	No change	,, K	Lower by 5 marks
,, E	Add 2½% to each mark	,, L	No change
,, F	Add 7½% to each mark	,, M	No change

104. *Limits for the various awards.*—The adjustments to the marks decided upon by the Chief Examiner were made in the office, and the thirty candidates examined by the six examiners were then treated as a hundred and eighty candidates examined by a single examiner and their marks arranged in order of merit. This list was sent to the Chief Examiner of the relevant Board,

who then fixed the limiting marks for the awards of Pass, Credit, and Distinction as follows[1] :—

Board I :—	Pass, 34 ;	Credit, 48 ;	Distinction, 72
Board II :—	Pass, 43 ;	Credit, 53 ;	Distinction, 73
Difference	9	5	1

105. *Adjusted marks.*—The adjusted marks and awards are set out in Tables 29 and 29A below :—

TABLE 29

TOTAL ADJUSTED MARKS

Board I		*Board II*	
Pass	34 marks	Pass	43 marks
Credit	48 marks	Credit	53 marks
Distinction	72 marks	Distinction	73 marks

Candidate	Examiner						Range	Examiner						Range	Extreme Range
	A	B	C	D	E	F		G	H	J	K	L	M		
1	48	44	51	51	46	49	7	56	56	52	50	53	53	6	12
2	76	73	71	66	76	78	12	83	85	74	76	80	70	15	19
3	67	63	66	66	70	72	9	74	73	66	70	76	71	10	13
4	47	39	41	47	37	44	10	36	48	37	48	48	46	12	12
5	33	32	25	35	24	25	11	27	27	29	28	29	34	7	11
6	19	17	16	22	16	14	8	26	26	22	19	19	19	7	12
7	48	47	42	37	31	33	17	21	38	49	45	38	40	28	28
8	62	68	74	66	57	71	17	69	80	65	65	75	76	15	23
9	68	64	62	60	62	63	8	70	74	67	72	78	68	11	18
10	14	12	16	12	6	9	10	5	14	3	7	7	11	11	13
11	23	28	21	17	23	24	11	27	33	32	29	36	33	9	19
12	47	53	47	48	38	47	15	47	50	53	52	46	51	7	15
13	43	42	46	42	41	41	5	57	53	55	53	57	54	4	16
14	49	51	53	49	53	47	6	63	56	57	49	54	52	14	16
15	66	65	71	59	73	68	14	77	70	71	66	80	71	14	21
16	53	47	46	48	49	49	7	49	59	54	61	61	53	12	15
17	35	34	34	32	34	31	4	36	39	50	43	45	46	14	19
18	23	37	20	26	24	23	17	31	30	35	30	35	30	5	17
19	13	15	9	10	9	9	6	18	21	14	15	16	14	7	12
20	36	42	34	33	31	31	11	35	40	41	48	40	44	13	17
21	58	63	61	59	65	62	7	73	76	61	70	72	78	17	20
22	65	65	63	59	69	70	11	80	78	73	76	82	84	11	25
23	61	54	58	55	56	57	7	64	66	60	65	68	68	8	14
24	57	52	56	54	59	52	7	55	68	58	61	69	65	14	17
25	45	44	49	45	46	43	6	57	63	54	61	61	58	9	20
26	45	45	41	45	45	37	8	49	56	51	53	49	51	7	19
27	82	74	75	71	72	71	11	80	76	71	78	77	76	9	11
28	45	47	39	41	44	44	8	51	55	50	58	51	53	8	19
29	53	52	51	52	53	55	4	61	71	56	70	59	62	15	20
30	88	81	85	70	88	95	25	94	89	88	87	90	86	8	25
Average	49·0	48·3	47·4	45·9	46·6	47·1	10·0	52·4	55·7	51·6	53·5	55·0	53·9	10·9	17·3

[1] It may be noted here that for Board I the maximum was 100, while for Board II it was 102.

106. *Ranges of Adjusted Marks.*—The extreme ranges for the two Boards are higher than for the French Boards. With Board I the range varies from 4 to 25; with Board II from 4 to 28.[1]

The average range for Board I is 10 marks; for Board II, 10.9.

TABLE 29A

AWARDS OF INDIVIDUAL EXAMINERS CALCULATED FROM ADJUSTED MARKS

	Board I						*Board II*					
	Pass 34 marks						Pass 43 marks					
	Credit 48 marks						Credit 53 marks					
	Distinction 72 marks						Distinction 73 marks					
Candidate	Examiner						Examiner					
	A	B	C	D	E	F	G	H	J	K	L	M
1	C	P	C	C	P	C	C	C	P	P	C	C
2	D	D	C	C	D	D	D	D	D	D	D	C
3	C	C	C	C	C	D	D	D	C	C	D	C
4	P	P	P	P	P	P	F	P	F	P	P	P
5	F	F	F	P	F	F	F	F	F	F	F	F
6	F	F	F	F	F	F	F	F	F	F	F	F
7	C	P	P	P	F	F	F	F	P	P	F	F
8	C	C	D	C	C	C	C	D	C	C	D	D
9	C	C	C	C	C	C	C	D	C	C	D	C
10	F	F	F	F	F	F	F	F	F	F	F	F
11	F	F	F	F	F	F	F	F	F	F	F	F
12	P	C	P	C	P	P	P	P	C	P	P	P
13	P	P	P	P	P	P	C	C	C	C	C	C
14	C	C	C	C	C	P	C	C	C	P	C	P
15	C	C	C	C	D	C	D	C	C	C	D	C
16	C	P	P	C	C	C	P	C	C	C	C	C
17	P	P	P	F	P	F	F	F	P	P	P	P
18	F	P	F	F	F	F	F	F	F	F	F	F
19	F	F	F	F	F	F	F	F	F	F	F	F
20	P	P	P	F	F	F	F	F	F	P	F	P
21	C	C	C	C	C	C	D	D	C	C	C	D
22	C	C	C	C	C	C	D	D	D	D	D	D
23	C	C	C	C	C	C	C	C	C	C	C	C
24	C	C	C	C	C	C	C	C	C	C	C	C
25	P	P	C	P	P	P	C	C	C	C	C	C
26	P	P	P	P	P	P	P	C	P	C	P	P
27	D	D	D	C	D	C	D	D	C	D	D	D
28	P	P	P	P	P	P	P	C	P	C	P	C
29	C	C	C	C	C	C	C	C	C	C	C	C
30	D	D	D	C	D	D	D	D	D	D	D	D

107. *Distribution of Awards.*—We shall return to the question of numerical marks later. We now deal with the awards of Pass, Credit, and Distinction.

Adopting the standards set out in para. 104 above, the

[1] The marks for Board II *before* adjustment showed one range of 33 marks.

distribution of awards by the several examiners of the two Boards is as follows :—

TABLE 30

	Board I						*Board II*					
Examiners	A	B	C	D	E	F	G	H	J	K	L	M
Failures	6	5	6	7	8	9	10	9	8	6	8	7
Passes	8	11	9	7	8	7	4	2	5	7	5	6
Credits	13	11	12	16	10	11	9	11	14	13	9	12
Distinctions	3	3	3	0	4	3	7	8	3	4	8	5
Total	30	30	30	30	30	30	30	30	30	30	30	30

Thus Examiner A of Board I awarded a Failure mark to six candidates, a Pass to eight, a Credit to thirteen, and Distinction to three candidates.

108. No two examiners agree entirely in the awards to the individual candidates. The degree of agreement obtained is indicated in Table 31 below :—

TABLE 31

NUMBER OF CASES IN WHICH THE EXAMINERS MAKE THE SAME AWARD

	Board I	*Board II*
	No. of Candidates	
Six examiners make the same award	14	13
Five examiners make the same award, one examiner differs	8	4
Four examiners make the same award, two agree on another award	6	9
Three examiners make the same award, three agree on another award	1	4
Three examiners make the same award, two agree on another, and one examiner differs	1	
	30	30

Thus the six examiners of Board I agree on the award to be given to fourteen out of the thirty candidates, and those of Board II on that to be given to thirteen candidates. There is complete agreement among the members of each Board separately in regard to rather under half the candidates.[1]

Those candidates about whose fate there is agreement amongst the examiners are not confined to one class.

[1] In French there was complete agreement on each Board in regard to slightly more than half the total number of candidates. See para. 72 above.

The fourteen candidates about whom agreement is reached by the examiners of Board I are distributed as follows: four candidates receive a Failure mark, four candidates receive a Pass, and six candidates receive a Credit.

Of the thirteen candidates in regard to whom agreement is reached by the examiners of Board II, six candidates receive a Failure mark, five candidates receive a Credit, and two candidates Distinction.

109. It will be seen from Table 30 that with Board I the number of Failures varies from 5 to 9, with an average of 7; with Board II, the number of Failures varies from 6 to 10, with an average of 8. The number of Passes for Board I varies from 7 to 11, with an average of 8; with Board II, it varies from 2 to 7, with an average of 5. In the matter of Credits, the Boards are more nearly alike: with Board I, the number varies from 10 to 16, with an average of 12, and with Board II from 9 to 14, with an average of 11. The number of Distinctions with Board I varies from 0 to 4, with an average of 3, while with Board II the number varies from 3 to 8, with an average of 6. The members of Board II award more Failure marks, fewer Pass marks, fewer Credit marks, and twice as many Distinctions as Board I.

110. Two striking cases of difference between the two Boards may be mentioned. Candidate No. 13 is awarded a Pass by every member of Board I and a Credit by every member of Board II, notwithstanding the fact that the limit for Credit with Board II is 5 marks higher than with Board I; and Candidate No. 22 is awarded a Credit by every member of Board I and Distinction by every member of Board II, though the limit for Distinction is a mark higher with the latter Board. Such differences must almost certainly be due to differences in the marking-schemes of the two Boards.

111. The differences in the awards of the two Boards depend in part on the differences in numerical marks, in part on the differences in the schemes of award (see para. 104 above). The average of the average marks of the different examiners of Board I is 47·4; for Board II the corresponding average is 53·7. The higher marks given by Board II compensate to some extent for their higher limits for Pass, Credit and Distinction, but there are, as we have seen, real differences between the general standards of the two Boards, apart from individual differences among examiners.

112. We have drawn attention in para. 106 above to the ranges of marks in Table 29.

Candidate No. 30 received the following marks from the examiners of Board I: 88, 81, 85, 70, 88, 95, showing a total range of 25 marks; with Board II his marks varied only from 86 to 94. Candidate No. 7 received the following marks from the examiners of Board II: 21, 38, 49, 45, 38, 40, showing a total range of 28 marks; with Board I, his marks varied from 31 to 48, a range of only 17 marks.

With both Boards the range goes as low as 4 marks. Candidate No. 17 received from Board I the following marks: 35, 34, 34, 32, 34, 31; while from Board II his marks varied from 36 to 50, a range of 14. For Candidate No. 29, the marks of Board I varied only from 51 to 55, while those of Board II varied from 56 to 71.

It is difficult to say why the paper of a candidate should be found easy to mark fairly concurrently by the members of one Board, while the members of the other Board disagree violently about his merits.

In the single case where the range of Board II was as low as 4 (the marks of Candidate No. 13 being 57, 53, 55, 53, 57, 54), the range for Board I was only 5 (the marks being 43, 42, 46, 42, 41, 41).

113. Table 32 below shows the distribution of the ranges in the case of each Board:—

TABLE 32

Range	No. of Candidates *Board I*	*Board II*
4	2	1
5	1	1
6	3	1
7	5	5
8	4	3
9	1	3
10	2	1
11	5	3
12	1	2
13	–	1
14	1	4
15	1	3
17	3	1
25	1	–
28	–	1
Total	30	30
Average Range	10·0	10·9

114. With Board I, the difference of marks between the limits for Pass and Credit is 14, and between those for Credit and Distinction, 24; there are six candidates out of thirty, one-fifth of the whole, for whom the range is 14 or more. With Board II, the difference between the limits for Pass and Credit is 10, and between those for Credit and Distinction 20; there are sixteen candidates with a range of 10 or more. Thus, in the case of Board II, the average range is greater than the span between Pass and Credit. [Cf. paras. 76 to 78 of the chapter on French School Certificate Scripts.]

115. The following Table shows how the averages of the examiners were affected by the adjustments of the Chief Examiners[1]:—

TABLE 33

Board I				*Board II*			
Examiner	Average Total Marks	Adjustment made by the Chief Examiner	Finally adjusted Average	Examiner	Average Total Marks	Adjustment made by the Chief Examiner	Finally adjusted Average
A	46·6	+2·4	49·0	G	52·4	0	52·4
B	51·3	−2·6	48·7	H	50·7	+5·0	55·7
C	45·0	+2·4	47·4	J	56·6	−5·0	51·6
D	45·9	0	45·9	K	58·5	−5·0	53·5
E	45·5	+1·1	46·6	L	55·0	0	55·0
F	43·8	+3·3	47·1	M	53·9	0	53·9

As we have pointed out in para. 81 of the chapter dealing with the investigation on French, no mere adjustment of averages would of itself remove the discrepancies between the distributions of awards by individual examiners.

As a result of the adjustments the difference between Examiners B and F of Board I is considerably reduced; but whereas Examiners C, D, E had nearly the same averages originally, the adjusted averages exhibit greater diversity, and it would seem that they should all have received the same treatment. In Board II, the adjustments appear to have raised Examiner H's marks too much, and to have lowered K's marks by too much, but on the whole the averages after adjustment are in better agreement than before.

116. It will be remembered that the candidates were requested to answer six questions out of eight. We subjoin in Table 34 below an analysis of the 2,160 marks allotted by the twelve examiners to the answers to these separate questions.

There is at this point a slight difference between this

[1] B's finally adjusted average, 48·7, differs slightly from the average shown in Table 29, as it was obtained by the subtraction of the final adjustment from the average of the original marks reduced to a percentage form, whereas the averages in Table 29 are the averages of the adjusted marks.

investigation and the parallel investigation on School Certificate French, since the candidates had in this examination a choice of questions not given in the French, and the number of answers to the various questions was therefore different.

TABLE 34

Board I

(1)		(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(*a*) Question		1	2	3	4	5	6	7	8
(*b*) No. of Answers		29	24	27	24	23	23	14	13
(*c*) Maximum Marks		16	16	16	16	16	17	16	16
Examiner									
	A	8·93	6·29	6·70	5·25	8·65	8·13	8·57	11·23
	B	9·28	7·83	7·56	5·67	8·87	8·87	9·57	11·38
(*d*) Average	C	8·48	7·08	6·96	4·87	8·00	7·61	8·79	10·46
Marks	D	8·24	5·75	6·93	6·00	8·26	9·26	8·64	10·46
	E	8·24	6·33	7·41	5·21	8·43	8·04	8·64	9·69
	F	8·83	7·62	6·52	4·75	7·13	7·61	7·71	10·15
Average of									
(*e*) Examiners' Averages		8·67	6·82	7·01	5·29	8·22	8·25	8·65	10·56
(*f*) % of Maximum Mark		54·2	42·6	43·8	33·1	51·4	48·5	54·1	66·0
(*g*) Highest Average Marks		9·28	7·83	7·56	6·00	8·87	9·26	9·57	11·38
(*h*) Lowest Average Marks		8·24	5·75	6·52	4·75	7·13	7·61	7·71	9·69
(*j*) Difference		1·04	2·08	1·04	1·25	1·74	1·65	1·86	1·69
(*k*) % of Maximum Mark		6·5	13·0	6·5	7·8	10·9	9·7	11·6	10·6
(*l*) Average Range		2·66	3·75	2·44	3·25	3·17	3·48	3·21	3·46
(*m*) % of Maximum Mark		16·6	23·4	15·2	20·3	19·8	20·5	20·1	21·6

Board II

(1)		(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
(*a*) Question		1	2	3	4	5	6	7	8
(*b*) No. of Answers		29	24	27	24	23	23	14	13
(*c*) Maximum Marks		17	17	17	17	17	17	17	17
Examiner									
	G	10·17	8·62	8·26	7·42	8·43	9·13	9·21	10·38
	H	9·41	7·33	7·41	8·04	7·57	9·09	9·64	12·31
(*d*) Average	J	11·14	8·83	8·44	7·37	8·57	11·39	10·29	11·92
Marks	K	10·93	8·92	9·15	8·71	9·09	10·57	10·79	12·69
	L	9·10	9·42	8·52	7·37	8·65	10·61	11·00	12·08
	M	9·83	8·58	8·48	7·96	9·13	8·83	9·50	12·31
Average of									
(*e*) Examiners' Averages		10·10	8·62	8·38	7·81	8·57	9·94	10·07	11·95
(*f*) % of Maximum Mark		59·4	50·7	49·3	45·9	50·4	58·5	59·2	70·3

Board II—continued

(1)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
(*g*) Highest Average Marks	11·14	9·42	9·15	8·71	9·13	11·39	11·00	12·69
(*h*) Lowest Average Marks	9·10	7·33	7·41	7·37	7·57	8·83	9·21	10·38
(*j*) Difference	2·04	2·09	1·74	1·34	1·56	2·56	1·79	2·31
(*k*) % of Maximum Mark	12·0	12·3	10·2	7·9	9·2	15·1	10·5	13·6
(*l*) Average Range	3·45	3·67	3·22	4·12	2·74	4·43	3·07	3·31
(*m*) % of Maximum Mark	20·3	21·6	18·9	24·2	16·1	26·1	18·1	19·5

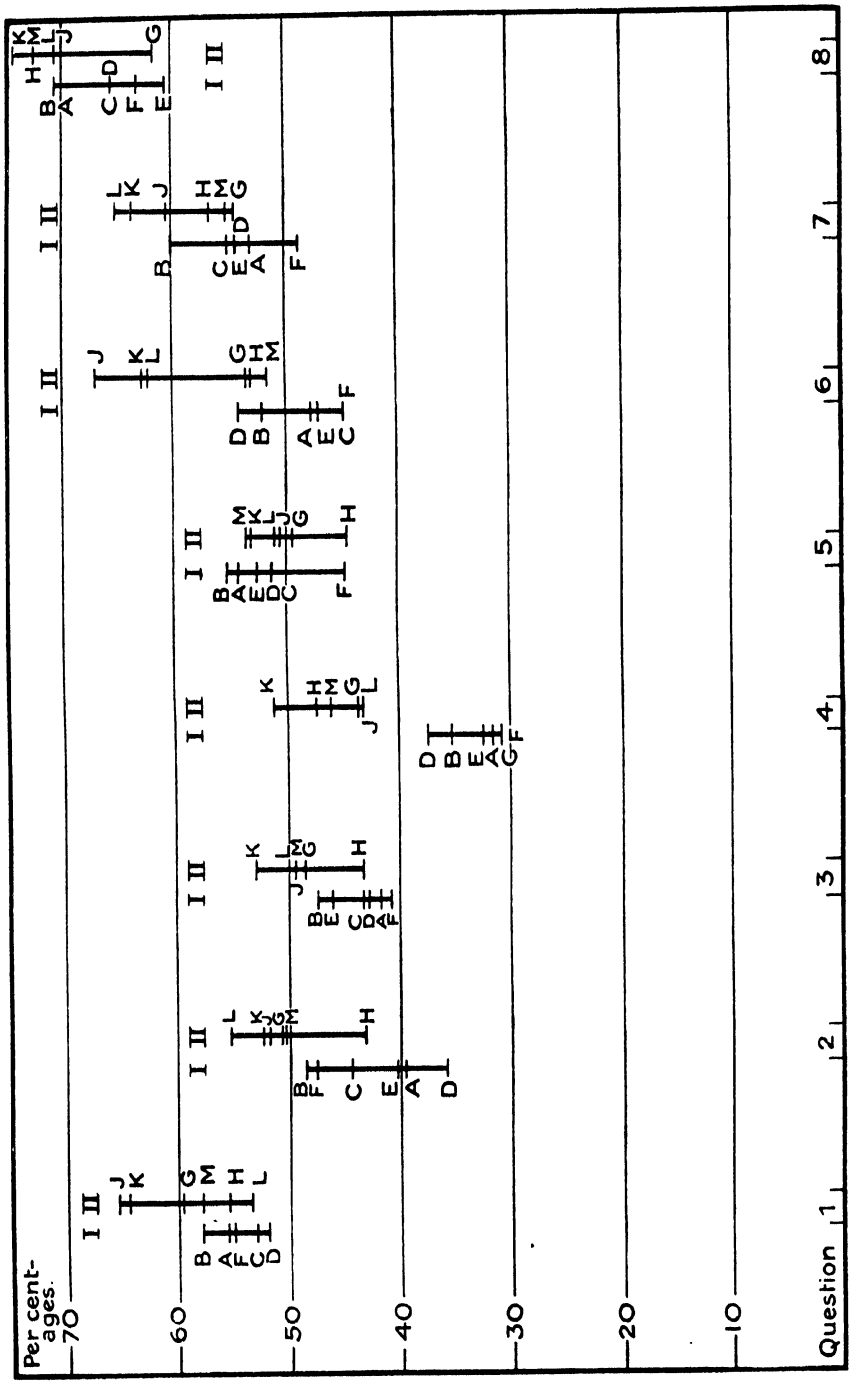
117. It is clear that in Board I, Examiner B marks all questions generously, and that Examiner F is on the whole the most severe; and again, in Board II, that Examiner K is the most generous and Examiner H the most severe.

We saw that the averages for the whole paper for Board II were higher than for Board I, and this difference again appears in the averages for the different questions. It appears most strikingly in the figures for Qn. 4, where the averages of the examiners for the two Boards do not even overlap, the averages of Board I ranging from 4·75 to 6·00, and those of Board II from 7·37 to 8·71. The average of the averages, expressed as a percentage of the maximum for this question, is 33 for Board I and 46 for Board II (see col. (5), row (*f*) and col. (13), row (*f*)). The question is a simple one, relating in part to chemical theory. It is only in regard to this point that we get anything comparable to certain remarkable differences which we found between the two French Boards (see para. 84 *et seq.*). Nevertheless, it may be said that the twenty-four candidates who selected Qn. 4 ran decidedly greater risks than those who avoided it. The two marking-schemes for the question differ decidedly. The diagram on p. 62 below shows graphically the average marks per question awarded by the several examiners, expressed as percentages of the maximum per question.

118. The diagram shows a general agreement between the two Boards as to the relative average values of the answers to the eight questions, though, as pointed out above, they disagree about the answers to Qn. 4. The unusually high average marks allotted to Qn. 8 call for attention. It was a question making demands on memory almost entirely, and not involving any kind of reasoning power.

119. The average range of marks for the different questions varies from 2·4 to 4·4, i.e. from about 15% to 26% of the

SCHOOL CERTIFICATE CHEMISTRY



maximum. The number of cases when the range is nil, i.e. when all the examiners of a Board agree on a mark for a candidate's answer to a question, is very small, as shown below :—

NUMBER OF CASES WHERE THE RANGE IS NIL

Question	*Board I*	*Board II*
1	0	1
2	2	0
3	4	2
4	1	1
5	0	1
6	0	0
7	1	1
8	0	0
Total	8	6

120. The results as a whole show the failure of the marking-schemes to secure either uniformity of marking or uniformity in the allotment of awards.

CHAPTER V

MARKING OF SCHOOL CERTIFICATE ENGLISH SCRIPTS

121. *Preliminary.*—The present chapter is based (1) on two papers by Mr. Charles Roberts and Professor H. V. A. Briscoe published in *The A.M.A.* (the organ of the Assistant Masters' Association) for December, 1931, and February, 1932, giving an account of an investigation carried out by the Durham University School Examination Board; (2) on further information, including the mark-sheets, kindly furnished to our Committee by the Durham Board.

122. *Character of the Examination Papers.*—There were two papers, Paper I was a two-hours paper, including both an essay (on a subject selected by the candidate from a list of eight) and a précis of a passage of 620 words from G. K. Chesterton, which the candidate was required to make one-third to one-quarter the length of the original. The passage for précis was given out at half-time.

Paper II was a three-hours paper, mainly dealing with set-books in prose and verse. The candidates were required to answer six questions, selected from twenty-four. The actual number of options was larger than these figures indicate, as many separate questions numbered singly included two or more options.

123. *Origin and Object of the Investigation.*—The original experiment arose from a difference of opinion as to the absolute merit of a group of candidates, and was designed specifically to investigate the extent of the differences which might occur among competent and experienced examiners in assessing the merits of the same candidates.

124. *Procedure.*—The whole of the English scripts from one school, forty-eight in number, were marked separately by seven examiners, A, B, C, D, E, F, and G, selected from the panels of four different School Certificate authorities, who had the reputation of being specially experienced and trusted examiners. Of

these, C, D, and E were ordinarily engaged by one authority, B and F by a second, and A and G by the third and fourth respectively.

The examiners all accepted the marking-scheme, including the limits for the different classes, of the Chief Examiner of the Durham Board.

The examiners did not meet or consult each other as to the marking ; no examiner had any knowledge of the others' marks ; but they were fully informed of the nature of the test in which they were taking part, and in particular each one was aware that one of the main objects of the investigation was the correct assessment of the class of the candidates.

125. The following Table shows the number of Failures, Passes, Credits, and Special Credits awarded by the different examiners :—

Examiner	*Failures*	*Passes*	*Credits*	*Special Credits*
A	1	16	27	4
B	0	2	34	12
C	7	30	11	0
D	0	9	36	3
E	5	16	27	0
F	2	7	37	2
G	19	12	17	0

The Table shows extraordinary discrepancies in the examiners' estimates of the candidates' capabilities.

126. The Credit mark is the most important mark in the scheme.[1] The following Table shows the different views of the examiners as to the number of candidates who reached and did not reach the Credit level.

Examiner	*Number of Awards below Credit*	*Number of Awards of Credit and over*
A	17	31
B	2	46
C	37	11
D	9	39
E	21	27
F	9	39
G	31	17

127. An inspection of the figures in greater detail shows that in the case of only *one* candidate out of the forty-eight were all seven examiners agreed as to the class in which he should be placed ; and there were only eight cases where six of the

[1] Because the exemption from Matriculation depends on the number of Credits obtained by a candidate.

examiners were in agreement. Examiner G "ploughs" nineteen candidates, while no other examiner "ploughs" more than seven and two "plough" none; Examiner B awards 12 Special Credits, while the other examiners award very few or none.

128. In all, twenty candidates were "ploughed" by one or more examiners. The following Table shows the awards of the other examiners to these candidates :—

"Ploughed" by	*No. of Candidates*	*Awards of the other Examiners*
4 Examiners	1	2 Passes ; 1 Credit
3 Examiners	2	4 Passes
		1 Pass ; 3 Credits
2 Examiners	7	4 Passes ; 1 Credit (in 2 cases)
		3 Passes ; 2 Credits (in 2 cases)
		2 Passes ; 3 Credits (in 2 cases)
		4 Credits ; 1 Special Credit
1 Examiner	10	5 Passes ; 1 Credit (in 2 cases)
		3 Passes ; 3 Credits (in 2 cases)
		2 Passes ; 4 Credits (in 4 cases)
		1 Pass ; 5 Credits
		5 Credits ; 1 Special Credit

129. Mr. Roberts and Professor Briscoe draw attention to certain extreme divergencies ; in Paper I (Essay and Précis)—

	Range of Marks
Candidate X was awarded 28, 32, 46, 56, 56, 58, 80, out of 100 by the seven examiners	52
Candidate Y was awarded 24, 42, 48, 60, 60, 64, 70, out of 100 by the seven examiners	46
Candidate Z was awarded 16, 36, 38, 44, 44, 46, 60, out of 100 by the seven examiners	44

On Paper I, nine candidates were awarded a Pass by all the examiners. Of the thirty-nine candidates who were not awarded a Pass by all the examiners, twenty-five were awarded a Credit, eight Special Credit, and three Distinction, by one or more examiners.

130. On Paper I, also, two of the examiners awarded between them Distinction to six candidates. The awards of the other examiners to these six candidates were as follows :—

No. of Candidate	*Awards of other Examiners*
1	Failure ; Pass ; Credit ; 3 Special Credits
2	Failure ; 4 Credits ; Distinction
3	2 Failures ; 4 Credits
4	2 Passes ; 4 Credits
5	Pass ; 3 Credits ; 2 Special Credits
6	4 Credits ; 2 Special Credits

131. In Paper II (Literature) the variations of award, though great, are somewhat less than in Paper I.

The marks of the candidates in regard to whom the divergencies were greatest, were as follows :—

Candidate	*Marks received from the seven Examiners (out of 100)*	*Range*
P	19, 41, 45, 46, 46, 49, 58	39
Q	37, 50, 52, 52, 54, 63, 71	34
R	38, 39, 45, 47, 53, 56, 70	32

132. Thirty-six of the forty-eight candidates were passed by all seven examiners in Paper II. Of the remainder, three were awarded a Failure mark by only one examiner, and eight were awarded a Failure mark by from two to four examiners ; but in all these cases the candidates were awarded from one to three Credits by other examiners. The nearest approach to unanimity was in the case of one candidate who was " ploughed " by six examiners, but was awarded a Credit by the seventh.

133. In Paper II, two of the examiners between them awarded Distinction to five candidates. The awards of the other examiners to these five candidates were as follows :—

No. of Candidate	*Awards of other Examiners*
1	Pass ; 4 Credits ; Special Credit
2	Pass ; 4 Credits ; Special Credit
3	2 Passes ; 4 Credits
4	6 Credits
5	3 Credits ; 3 Special Credits

134. The number of awards of the different examiners on Papers I and II separately, divided into the categories " Below Credit " and " Credit and Over," are shown below[1] :—

	PAPER I (*Essay and Précis*) *Number of Awards*		PAPER II (*Literature*) *Number of Awards*	
Examiner	*Below Credit*	*Credit and Over*	*Below Credit*	*Credit and Over*
A	22	26	14	34
B	17	31	2	46
C	41	7	31	17
D	26	22	8	40
E	28	20	23	25
F	27	21	8	40
G	46	2	25	23

[1] It is interesting to note the general opinion of the examiners that the standard in Set Books was much higher than in Précis and Essay. See the article on English at the School Certificate Examination by one of the present writers in the " Essays on Examinations " published by the Committee.

Further details of the Durham investigation are given by Professor Briscoe in *The A.M.A.* for March, 1932 (p. 78), and by Mr. C. Roberts in the *Journal of Education* for April, 1932 (p. 225).

CHAPTER VI

SPECIAL PLACE EXAMINATION (I): MARKING OF ARITHMETIC AND ENGLISH SCRIPTS

SECTION I. INTRODUCTORY

135. *Preliminary.*—The name "Special Place Examination" is a title commonly given to those competitive examinations held by Local Education Authorities on the results of which pupils of elementary schools in this country are either (a) awarded scholarships enabling them to proceed to a public secondary school under the same authority, (b) given admission to a school for boys or girls up to the age of 15, known as a "Selective Central" or simply a "Central" school, or (c) retained until the "school-leaving age" in the upper forms of an elementary school.[1]

136. It is obvious that these competitive examinations are of great importance, both from the point of view of the individual child and of the general community. (See the paper on *The Special Place Examination* by Dr. Ballard in the "Essays on Examinations" published by the Committee.) It is of the utmost importance to ensure that they should fulfil their purpose of selecting the pupils best fitted to profit by higher education, in other words that their "validity" as tests should be high. But, clearly, no test can be a "valid test" unless it yields consistent results in the hands of different examiners, i.e. unless its "reliability" (to use the word generally employed by educational psychologists), or, as we should prefer to say, its "consistency," is high. In the present series of investigations we are dealing only with the question of consistency. To say that the "consistency" is low, is another way of saying that the element of chance in the examination is great. The general question of validity must be reserved for further consideration.

[1] Other titles given to the examination are "Free Place Examination," "Junior Scholarship Examination" and "Eleven Plus Examination."

137. *Character of the Examination Papers.*—The examination of which we investigated the scripts included two subjects, Arithmetic and English, and each paper was divided into two sections, A and B.

Arithmetic.—Section A consisted of twenty short sums, for which 35 minutes was allowed. Section B consisted of five questions of which four were problems, for which 40 minutes was allowed.

English.—Section A of the English paper consisted in the telling of a story (40 minutes allowed). Section B consisted of a series of detailed questions mainly relating to the meaning and use of words (35 minutes allowed).

Further details of these papers will be given at a later stage in this chapter.

138. *Precise Object of the Investigation.*—The object of the investigation was to ascertain the consistency in the marking of the scripts by ten examiners for each subject, all experienced in marking scripts in that subject at Special Place examinations.

Some of the examiners were chosen from the panel of the authority by whom the scripts were furnished, but they had not acted at this particular examination. Other examiners were selected from the panels for Special Place examinations of three other authorities.

139. *Procedure : Selection of Scripts.*—We were furnished by the authority concerned with over three hundred scripts selected from a very much larger number. From these, one hundred and fifty scripts were selected for the final investigation, including a large proportion of those which obtained the highest marks at the original examination, and the rest were reserved for use as trial scripts.

Every mark on each of the scripts indicating either its origin or the marks originally allotted to it was obliterated.

140. Fifty trial scripts were sent to each of the examiners for preliminary marking, with a draft marking-scheme for criticism ; and it was only after considerable correspondence with the examiners individually and careful consideration of every point raised that the marking-schemes in Arithmetic and in English were finally revised and settled. The schemes in both subjects were discussed with Dr. Ballard ; and those in Arithmetic were also discussed with another examiner with long experience of this kind of examination. Our aim was in both subjects to arrive at schemes which would be free from any kind of ambiguity. We felt that after the ten examiners in each subject had each marked fifty trial scripts and raised every

point of difficulty that occurred to them, we should have reason for thinking that the ground had been amply covered, both for English and for Arithmetic.[1]

141. The sets of one hundred and fifty scripts referred to in para. 139 above were circulated in original to the ten examiners concerned in each case, together with the revised marking-scheme and appropriate marking-sheets. The circulation of the scripts naturally took a considerable period. The amount of time required for the marking of the scripts was settled by the examiner himself in each case. Any element of error due to working under pressure was therefore eliminated.

142. To print in detail the marks awarded would occupy an undue amount of space. But in order to give the reader a preliminary impression of the results obtained, we print in Appendix II (pp. 114–116 below), the marks for Sections A and B in English, and for Sections A and B in Arithmetic, awarded to every *fifth* candidate on each list, by the ten examiners concerned.

To avoid misunderstanding it should be added that although we have both in Arithmetic and in English designated the examiners by the letters, A, B, C, D, E, F, G, H, J, K, the two sets of examiners for English and for Arithmetic were entirely distinct.

Section II. The Combined Results of the Examinations in Arithmetic and English

143. In a real examination the fate of each candidate would primarily depend on marks allotted by one examiner in Arithmetic and one in English. Since in our investigation there were ten examiners in each subject, there were one hundred possible pairs of examiners. But to investigate the results likely to be obtained it was sufficient to choose any pairs at random, and, as the distinguishing letters by which the examiners were designated were given in a random manner, for the sake of convenience Examiner A in Arithmetic was grouped with Examiner A in English, Examiner B in Arithmetic with Examiner B in English, and so on; and the ten couples so obtained were designated Couples A, B, C . . . K.

We thus obtained ten sets of total marks such as might have been obtained at the original examination.

[1] Both for Arithmetic and for English the total number of trial scripts used was one hundred; one set of fifty went to five examiners out of the ten concerned, and another set to the other five.

The total possible mark for any candidate was 200 (100 for each subject).

144. There were no cases where the marks awarded to a candidate by the ten couples were the same ; in many cases they were strikingly different.

Table 35 below shows the lack of concordance between the awards of the ten couples of examiners :—

TABLE 35

	Number of Candidates
Assigned different marks by all ten couples	38
Assigned the same mark by only two couples	64
Assigned the same mark by only three couples	4
Assigned the same mark by only four couples	2
Assigned one mark by two couples, another mark by two other couples	26
Assigned one mark by two couples, another mark by three other couples	8
Assigned one mark by two couples, another mark by two other couples, and a third mark by two other couples	6
Assigned one mark by two couples, another mark by two other couples, and a third mark by three other couples	2
	150

A single example will illustrate the meaning of this Table. Candidate No. 1 received the following marks from the ten couples of examiners : 139, 124, 124, 136, 110, 130, 107, 119, 109, 105. This was one of the cases in which two marks were the same. The range (i.e. the difference between the highest and lowest mark) is here quite large, 34 marks.

145. The distribution of the ranges for the different candidates is shown in Table 36 below :—

TABLE 36

Range	12–16	17–21	22–26	27–31	32–36	37–41	42–46	47–51	52–56	57–61	62–66
Number of Candidates	3	12	22	27	43	22	12	4	2	2	1

Total 150

The smallest range is 12, in the case of Candidate No. 38, who received from the different couples of examiners 140, 136, 139, 140, 140, 141, 129, 132, 134, and 129 marks. The largest range is 63, for Candidate No. 53, who received 132, 129, 146, 131, 168, 125, 120, 128, 105, and 124 marks.

The average range for the whole group is 33 marks, that is 16½ marks out of 100.

146. This range must be regarded as considerable in view of the fact that the examiners were all experienced in this type of work, and that they were marking according to carefully drawn up marking-schemes. It is also important from the point of view of the candidates, since the examination is one for scholarships, in which a few marks might make the difference between success and failure, and where, consequently, success or failure might depend on the particular couple of examiners who marked the scripts.

147. The average marks awarded by the ten couples of examiners were as follows: A, 123; B, 125; C, 130; D, 135; E, 141; F, 130; G, 116; H, 126; J, 120; K, 115. Thus, Couple E marks high, while Couples G and K mark low. But a difference of 26 marks between the average of Couple E and that of Couple K is very high considering the circumstances of the examination.

148. Where there are many assistant-examiners the Chief Examiner scrutinises the marks and makes adjustments for different standards of marking, as shown by their averages. The distributions of the marks are also sometimes reduced to a standard. No such adjustment would alter the order of the candidates in the batch assigned to a single assistant-examiner; yet it is this question of order which is of fundamental importance in a scholarship examination.

149. Let us therefore consider the order of merit in which the candidates are placed by the different couples of examiners. Table 37 below shows the different candidates (out of the one hundred and fifty) who were assigned the first ten places on the list.

TABLE 37
Couples of Examiners

A	B	C	D	E	F	G	H	J	K
Candidates, designated by their roll-numbers									
81	81	75	29	75	29	135	29	68	81
29	65	64	47	81	3	29	81	29	135
75	3	81	30	122	65	64	30	135	75
135	79	65	68	126	135	3	75	103	47
47	75	29	79	29	9	9	68	75	64
64	133	79	122	68	122	68	122	81	3
55	126	122	3	69	75	75	135	149	79
68	9	9	64	103	131	103	9	65	87
88	29	68	103	135	103	47	65	147	55
65	135	69	55	85	149	82	133	64	122
	148					149			

Thus Candidate No. 81 is placed first by three couples of examiners, second by two, and by others third and fifth, and he is placed just below the first ten by the rest. There is no candidate who is placed in the first ten by all the ten couples.

150. Let us suppose for a moment that fifty scholarships were to be awarded as the result of the examination.[1] There are seventy-three candidates who are placed among the first fifty by one or other of the couples of examiners. In the lists of Couples A, D and E, the fiftieth place is occupied by more than one candidate, and we should have to consider fifty-two candidates on each of their lists. The following Table shows the result of such a scrutiny :—

TABLE 38

33	candidates	are	returned	in	the first 50	by all 10 couples
8	,,	,,	,,	,,	,,	,, 9 couples
4	,,	,,	,,	,,	,,	,, 8 ,,
4	,,	,,	,,	,,	,,	,, 7 ,,
1	candidate	is	,,	,,	,,	,, 5 ,,
1	,,	,,	,,	,,	,,	,, 4 ,,
3	candidates	are	,,	,,	,,	,, 3 ,,
7	,,	,,	,,	,,	,,	,, 2 ,,
12	,,	,,	,,	,,	,,	,, only 1 couple
73						

151. Thus (if fifty scholarships were awarded), thirty-three candidates would get scholarships whichever couple of examiners marked their scripts ; but the fate of the other seventeen candidates would depend on the chances of being assigned to particular couples, the chance being greater for some candidates than for others.

The thirty-three candidates about whom there is agreement are Nos. 3, 9, 15, 18, 29, 30, 47, 55, 64, 65, 67, 68, 75, 76, 79, 81, 82, 87, 95, 103, 122, 123, 124, 126, 131, 133, 134, 135, 138, 147, 148, 149, 150. The eight candidates about whom nine couples agree that they should be in the first fifty are Nos. 49, 69, 77, 83, 85, 88, 107, 140. The four candidates about whom eight couples agree are Nos. 27, 33, 36, 46. The four about whom seven couples agree are Nos. 7, 23, 93, 109. The one candidate about whom five agree is No. 129.

152. Let us consider the fifty candidates above-mentioned,

[1] This number is not the number awarded at the examination which yielded our scripts.

about each of whom there is agreement amongst at least five couples (not necessarily the same five in each case). The views of all the various couples with regard to this list are expressed in the following scheme :—

TABLE 39

Couples of Examiners						
A would add Nos. 38 and 42 to the first 52 candidates (by bracketing)						
B	would exclude Nos.		7, 109, 129	but would include Nos.		8, 73 and 50 in the first 50 candidates
C	,,	,,	33, 77, 129	,,	,,	53, 56, 128 in the first 50 candidates
D	,,	,,	46, 93, 140, 109	,,	,,	8, 22, 105, 116, 121, 128 in the first 52 candidates
E	,,	,,	46, 49, 88	,,	,,	8, 56, 97, 105, 110 in the first 52 candidates
F	,,	,,	23, 85, 107	,,	,,	2, 5, 105 in the first 50 candidates
G	,,	,,	23, 36, 69, 93	,,	,,	98, 110, 128, 130 in the first 50 candidates
H	,,	,,	23, 83, 129	,,	,,	59, 97, 98 in the first 50 candidates
J	,,	,,	7, 27, 36, 93	,,	,,	38, 72, 115, 116 in the first 50 candidates
K	,,	,,	27, 33, 109, 129	,,	,,	38, 40, 48, 57 in the first 50 candidates

The differences are not great. Couple A agrees most nearly with the list of fifty selected in the way described. The other couples agree remarkably in their method of disagreeing, since each couple drops three or four candidates and replaces them by another three or four ; but with each couple the selection changes of those dropped and those added.

153. Let us now consider the candidates at the bottom of the general list instead of those at the top, and ascertain to which candidates the various couples would assign the last fifty places. Owing to bracketing at the bottom we have to consider 52 candidates in the lists of Couples A, C, D, E, G, J, and K, and 51 in H's list.

We shall have to consider altogether the claims of ninety-five candidates for these lowly places, as shown below :—

TABLE 40

Couples of Examiners					No. of Candidates
All 10	agree	in	placing	in the last fifty	4, 12, 42, 45, 71, 84, 90, 112, 143 — 9
9	,,	,,	,,		14, 17, 26, 31, 34, 35, 39, 43, 44, 52, 74, 89, 96, 117, 119 — 15
8	,,	,,	,,		19, 24, 61, 102, 106, 114, 145, 146 — 8
7	,,	,,	,,		13, 25, 41, 59, 80, 94, 99, 104, 111 — 9
6	,,	,,	,,		66, 97, 120 — 3
5	,,	,,	,,		16, 60, 63, 73, 78, 91, 136 — 7
4	,,	,,	,,		. . . . — 12
3	,,	,,	,,		. . . . — 11
2	,,	,,	,,		. . . . — 8
1 couple only place in the last fifty					. . . — 13
					95

154. Of the ninety-five candidates included by one or more couples in the lowest third of the whole group, there is complete agreement in regard to only nine candidates, as contrasted to the complete agreement in regard to thirty-three candidates in the highest third. If, proceeding as before, we take the list of fifty-one candidates about whom there is agreement among five or more couples of examiners (not the same couples for each candidate), we find that the various couples would desire the following changes to be made in it :—

TABLE 41

Couple										
A	would	exclude	10	but	would	include	11	others	in the last	52 places
B	,,	,,	10	,,	,,	9	,,	,,	50	,,
C	,,	,,	8	,,	,,	9	,,	,,	52	,,
D	,,	,,	8	,,	,,	9	,,	,,	52	,,
E	,,	,,	15	,,	,,	16	,,	,,	52	,,
F	,,	,,	8	,,	,,	7	,,	,,	50	,,
G	,,	,,	13	,,	,,	14	,,	,,	52	,,
H	,,	,,	11	,,	,,	11	,,	,,	51	,,
J	,,	,,	11	,,	,,	12	,,	,,	52	,,
K	,,	,,	11	,,	,,	12	,,	,,	52	,,

155. There is obviously less precision in assigning a right order to the weaker than to the stronger candidates.[1]

And, considering the results as a whole, it is clear that even when very great care is taken in drawing up and discussing marking-schemes, on the basis of a preliminary marking of trial scripts, at an elementary examination of this kind, anything like the machine-like precision which is sometimes supposed to be attained is a figment of the imagination, and that with all the safeguards introduced there is much room for the personal equation of the examiners to enter.

156. We shall now examine the results in Arithmetic and in English separately in order to ascertain how far the discrepancies depend on the subject-matter of the papers.

Section III. Arithmetic

157. As stated above, the paper set consisted of two parts, A and B. The maximum mark for Part A was 40, and for Part B, 60. Part A consisted of 20 short questions, each carrying 2 marks, while Part B had 5 questions, each carrying 15 marks, but the marks of only the best four answers were counted.

One hundred and fifty scripts were marked by ten examiners, after the agreed marking-scheme had been established in the manner described in paras. 139 and 140 above.

158. The results are remarkable. In the case of only a single candidate was there complete agreement among the examiners—Candidate No. 96 with 41 marks. The greatest difference in the marks allotted to a single candidate was 39, Candidate No. 116 being awarded 65, 65, 65, 65, 77, 85, 58, 65, 89, 50 by the different examiners.

159. It is to be remembered that each candidate has 10 marks assigned to him by the different examiners (though some of these marks may be the same), and, in order to give a rough conspectus of the distribution of the marks, we have divided the candidates into groups according to the highest and lowest marks obtained by candidates in a group.

[1] It must be remembered that no very weak candidates are included in our list.

TABLE 42[1]

Group	Limit of highest marks in the Group	Limit of lowest marks in the Group	No. of candidates in the Group	
I	90 to 100	90 to 100	5	
II	90 ,, 100	80 ,, 89	12	
III	90 ,, 100	70 ,, 79	8	
IV	90 ,, 100	60 ,, 69	3	
V	80 ,, 89	80 ,, 89	1	
VI	80 ,, 89	70 ,, 79	12	
VII	80 ,, 89	60 ,, 69	14	
VIII	80 ,, 89	50 ,, 59	7	
IX	80 ,, 89	40 ,, 49	1	Total 63
X	70 ,, 79	70 ,, 79	2	
XI	70 ,, 79	60 ,, 69	11	
XII	70 ,, 79	50 ,, 59	10	
XIII	70 ,, 79	40 ,, 49	2	
XIV	60 ,, 69	60 ,, 69	1	
XV	60 ,, 69	50 ,, 59	23	
XVI	60 ,, 69	40 ,, 49	12	
XVII	60 ,, 69	30 ,, 39	1	
XVIII	50 ,, 59	50 ,, 59	7	
XIX	50 ,, 59	40 ,, 49	7	
XX	50 ,, 59	30 ,, 39	2	
XXI	40 ,, 49	40 ,, 49	5	
XXII	40 ,, 49	30 ,, 39	3	
XXIII	40 ,, 49	20 ,, 29	1	
			150	

160. Thus there are sixty-three candidates who get 80 or more marks from at least one examiner, and of these eighteen get 80 or more from all examiners. If we regarded 80 as a high mark intended to indicate scholarship level, this would mean that there is complete agreement in regard to only eighteen out of the sixty-three possibles.

161. The following Table shows the distribution of the ranges of marks. The number in the second row shows the number of candidates whose marks showed the range corresponding in the first row ; thus the number of candidates whose marks from the different examiners showed a range of 16 was ten ; and the number of those for whom the range was 20 was eight.

[1] It may perhaps make this Table clearer if its construction is explained by means of a single example. Let us suppose that to a certain candidate AB the highest mark assigned by any examiner is 57 and the lowest assigned by any examiner is 45. He can then only come into Group XIX. If his highest mark had been 61 but his lowest mark had been 45, he would be placed in Group XVI, and if the highest mark had been 73, he would be placed in Group XIII.

TABLE 43

ARITHMETIC

Range	0	3	4	5	6	7	8	9	10	11	12
Number of Candidates	1	1	5	8	1	6	14	3	3	16	12
					(*continued*)						
Range	13	14	15	16	17	18	19	20	21	22	23
Number of Candidates	5	3	10	10	2	6	5	8	4	3	6
					(*continued*)						
Range	24	25	26	27	28	29	32	38	39		
Number of Candidates	1	4	1	5	2	2	1	1	1	Total 150	

162. The average range is 14·7 marks (out of 100). There are only 39 cases in which the examiners agree within 10 marks. There are also 39 cases in which the maximum difference amounts to 20 marks or more.

163. We shall now consider how far the discrepancies are due to the two parts of the paper.

164. Part A, it will be remembered, consisted of straightforward sums, and it was expected that the marks of the ten examiners would be identical. The maximum mark was 40 and the greatest difference between any two examiners was 8 marks. As to ninety-eight candidates out of the one hundred and fifty, all the examiners were agreed. Table 44 below shows the measure of agreement reached.

TABLE 44

ARITHMETIC PAPER, PART A

		Number of candidates
10 examiners agree		98
9 „ „	1 differs	27
8 „ „	2 agree together	9
8 „ „	2 differ	4
7 „ „	3 agree together	5
7 „ „	3 differ	1
6 „ „	4 agree together	3
6 „ „	2 agree, 2 agree	1
5 „ „	5 agree together	1
4 „ „	4 agree, 2 agree	1
		150

It will be seen from para. 165 that to forty-five candidates two different marks were allotted ; to seven candidates three different marks were allotted.[1]

165. The distribution of the ranges was as follows :—

Range	Candidates
0	98
2	45
4	4
6	1
8	2
	150

The average range was 0·85, corresponding to 2·1% of the maximum. Since the average range for the whole paper was 14·7%, it follows that the greatest discrepancies are due to Part B.

166. Part B contained 5 questions, of which 4 were problems. The maximum was 60. Fifteen marks were allotted to each question, but only the four best answers were counted.

167. The measure of agreement between the different examiners in regard to the several questions is shown in Table 45 below.

[1] The following illustration shows in what way the differences have arisen :—

(i) One question read as follows :—

Divide 7·83 by 0·09

and the instructions in the marking-scheme were to mark as correct only 87, or 87·, or 87 times. One candidate gave as his answer 87·00. Of the ten examiners, six gave 2 marks (the maximum) and four gave 0.

(ii) Another question was :—

To find the value in decimals of $\frac{4{\cdot}125}{13{\cdot}75}$

and the instructions in the marking-scheme were to mark as correct the following answers only :—

0·3, or ·3, or ·30, but *not* $\frac{3}{10}$ or ·003 or ·01

or variations of this.

One candidate wrote $\frac{3}{10}$ = ·30f [*sic*]. Five examiners gave 2 marks, and five gave no marks. Another candidate wrote ·3, but the decimal point was written very faintly, and this was presumably the reason why only two examiners gave 2 marks, while eight gave none.

TABLE 45

ARITHMETIC PAPER, PART B

Question						1	2	3	4	5
10 examiners agree as to the following number of candidates						Number of candidates				
						98	90	35	20	61
9	,,	,,	,,	,,	,,	26	21	58	15	25
8	,,	,,	,,	,,	,,	9	11	16	12	16
7	,,	,,	,,	,,	,,	5	6	5	16	9
6	,,	,,	,,	,,	,,	7	4	19	20	10
5	,,	,,	,,	,,	,,	4	3	6	34	6
4	,,	,,	,,	,,	,,	1	1	5	18	2
3	,,	,,	,,	,,	,,				3	
						150	136	144	138	129

Thus Qns. 1 and 2 are comparatively easy to mark, and the examiners are in complete agreement in regard to about two-thirds of the candidates ; Qn. 5 is less easy ; in regard to Qns. 3 and 4 the discrepancies are very considerable.

168. A complete analysis of the marks awarded for Qn. 4 is given in Table 46 below :—

TABLE 46

	Marks						Corresponding Number of Candidates	
	15	12	8	7	4	0		
Examiners	9	1					1	
,,	9		1				1	
,,	9			1			2	
,,	9				1		2	
,,	9					1	1	
,,					9	1	3	
,,		1			9		1	
,,					1	9	4	
							——	15
,,	8	2					1	
,,	8	1		1			1	
,,	8	1			1		1	
,,	8				1	1	1	
,,				8	2		1	
,,			2		8		1	
,,					8	2	2	
,,			1		8	1	1	
,,					2	8	2	
,,		1			1	8	1	
							——	12
					Carried forward			27

TABLE 46—*continued*

	Marks						Corresponding Number of Candidates	
	15	12	8	7	4	0		
					Brought forward			27
Examiners	7	1			1	1	1	
,,	7			2	1		1	
,,	7			3			1	
,,				7	3		1	
,,	1			7	1	1	2	
,,			1	7	1	1	1	
,,					7	3	3	
,,			1	2	7		1	
,,			1		7	2	1	
,,				1	7	2	2	
,,					3	7	2	
								16
,,	6	1		3			1	
,,	6			1	1	2	1	
,,	6	1	1	1	1		1	
,,			6		2	2	1	
,,				6	4		3	
,,				6	2	2	1	
,,	1	1	1	6	1		1	
,,	2			6	2		2	
,,	2			6	1	1	1	
,,		2		6	2		1	
,,				4	6		1	
,,				3	6	1	1	
,,					6	4	1	
,,			4		6		1	
,,			1		3	6	1	
,,					4	6	2	
								20
,,	5	2	1	1	1		1	
,,	5	1		1	3		1	
,,	5	2		1	2		1	
,,	1	3	5		1		1	
,,	2	2	5			1	1	
,,			5		4	1	1	
,,				5	4	1	2	
,,	1			5	2	2	2	
,,				5	3	2	2	
,,				5	5		1	
,,	2			5	3		1	
,,		1		5	3	1	1	
,,	3			5	1	1	1	
,,		1	1	5	2	1	1	
,,			2	5	3		1	
,,	4	1		5			1	
,,	4		1	5			1	
,,	3		1	5	1		1	
					Carried forward			63

TABLE 46—*continued*

	Marks						Corresponding Number of Candidates	
	15	12	8	7	4	0		
					Brought forward			63
Examiners					5	5	5	
,,	1		1	2	5	1	1	
,,				1	5	4	1	
,,	1	2			5	2	1	
,,				4	5	1	1	
,,	2		1	1	1	5	1	
,,				2	3	5	1	
,,		1	1		3	5	1	
,,		2		1	2	5	1	
							——	34
,,	4	3		1	2		1	
,,	4			2	1	3	1	
,,	4			4	1	1	1	
,,	4			4	2		1	
,,	1	1		4	3	1	1	
,,	3	2		4	1		1	
,,	2		1	4	3		1	
,,	3			4	2	1	1	
,,	1	1		4	4		1	
,,		2	1	4	3		1	
,,	3	1		2	4		2	
,,				3	4	3	1	
,,		1	2	2	4	1	1	
,,	2	1	1	2	4		1	
,,	1			3	4	2	1	
,,			3		3	4	1	
,,	3			1	2	4	1	
							——	18
,,	3	2	1	1	2	1	1	
,,	1	3	3	1	1	1	1	
,,	1	3	2	2	2		1	
							——	3
								118
			10 Examiners agree					20
								138

169. The following illustration shows how this Table is to be read : The first candidate shown on the list gets 15 marks from nine examiners and 12 marks from one examiner ; the last candidate shown on the list gets 15 marks from one examiner, 12 from three examiners, 8 from two, 7 from two, 4 from two examiners.

170. Twenty-two candidates get marks ranging from 0 to 15 inclusive from one or more examiners. The marks awarded to thirty-three candidates differ by no more than 3 or 4. In the case of twenty candidates the examiners' marks are exactly the same.

171. The problem type of question used is presumably set with the idea of bringing out the candidates' powers of reasoning more fully than is possible with the questions in Part A or with Qn. 1 in Part B, and real distinction should be exposed in the answers.

But the elaborate precautions taken to secure that all examiners shall mark in identical fashion every answer, complete or partial, have failed, in spite of an assurance from each of them that the revised marking-scheme left no room for doubts. With problems of this kind it would seem, therefore, that the mark awarded will in many cases depend to a considerable extent on the personal equation of the examiner.

Section IV. English Essay Marks

172. As stated above, the English paper consisted of two Parts, A and B, of which A was an essay-paper. The maximum for each part was 50 marks. Further details of the marking-scheme are given in Appendix I to this chapter, pp. 112–113 below.

The scheme which we used for marking the essay was based on schemes used in this country and in the United States, in which marks are allocated separately for the different " elements " in the composition. It must not be assumed that we regarded the analysis into elements used in this particular scheme with any special favour. We simply wished to ascertain what results it would yield.

173. Each examiner was required to allot one of the following five marks, 0, 1, 3, 5, 7, in respect of each of the following seven elements of the composition—Vocabulary, Accuracy, Craftsmanship, Consistency, Completeness, Substance, and Quality.[1] Each of the elements carried an equal weight in the assessment of the total marks ; and an odd mark was left over to be allotted at the discretion of the examiners.

[1] In our second investigation on Special Place English scripts we adopted for the marking of half these scripts a scheme on somewhat the same principle, but differing in important details (see para. 251 *et seq.* below).

174. The discrepancies between different examiners in allotting these marks were considerable. Table 47 below[1] gives a conspectus of the agreement between the different examiners :—

TABLE 47

NUMBER OF CASES OF AGREEMENT AMONGST EXAMINERS

Examiners agreeing	Vocabulary	Accuracy	Crafts-manship	Con-sistency	Com-pleteness	Substance	Quality
10		1			1	1	
9, 1	2	5	3	1	2	4	2
8, 2	8	6	8	3	4	7	6
8, 1, 1	7	7	7	4	5	2	4
7, 3	9	13	8	7	6	11	11
7, 2, 1	6	13	13	8	8	17	9
7, 1, 1, 1	5	1		1	1	2	1
6, 4	8	10	13	6	8	15	8
6, 3, 1	14	16	20	15	20	15	20
6, 2, 2	5	12	9	3	11	7	11
6, 2, 1, 1	5	7	5	9	1	3	4
5, 5	6	6	7	2	6	5	6
5, 4, 1	17	15	18	14	24	24	21
5, 3, 2	18	12	15	18	16	14	14
5, 3, 1, 1	9	6	4	10	2	2	4
5, 2, 2, 1	1	4	4	9	3	3	4
5, 2, 1, 1, 1				4			
4, 4, 2	11	3	3	4	10	7	6
4, 4, 1, 1	2	2	5	2	3	4	4
4, 3, 3	6	5	3	9	7	5	7
4, 3, 2, 1	6	3	3	14	9	2	7
4, 2, 2, 2	2			2	2		
4, 3, 1, 1, 1		1		1	1		
3, 3, 3, 1	3	1	1	1			
3, 3, 2, 1, 1		1					
3, 3, 2, 2			1	3			1
	150	150	150	150	150	150	150

In the above Table, " 6, 3, 1 " means that six examiners gave the same mark, three gave another mark, and one examiner gave a different mark. Or again, " 5, 2, 1, 1, 1 ", " 4, 3, 1, 1, 1 " and " 3, 3, 2, 1, 1 " mean that all the possible marks, 0, 1, 3, 5, 7, were awarded to a candidate for some element by one or more examiners. There were only three occasions when the examiners all agreed in their judgment of an element.

175. Table 48 below shows the number of candidates in regard to whom agreement had been reached in respect of the different

[1] Compare Table 89.

elements, Vocabulary, Accuracy, etc., by ten, nine, eight, etc., examiners :—

TABLE 48

Number of cases of agreement

	Vocabulary	Accuracy	Crafts-manship	Con-sistency	Com-pleteness	Substance	Quality
10 agree	0	1	0	0	1	1	0
9 ,,	2	5	3	1	2	4	2
8 ,,	15	13	15	7	9	9	10
7 ,,	20	27	21	16	15	30	21
6 ,,	32	45	47	33	40	40	43
5 ,,	51	43	48	57	51	48	49
Rest	30	16	16	36	32	18	25
	150	150	150	150	150	150	150

176. Thus in respect of Vocabulary no candidate out of the hundred and fifty receives the same mark from all examiners ; only two candidates receive the same mark from nine examiners ; only fifteen candidates receive the same mark from eight examiners, and so on.

177. There is nothing like unanimity among the examiners in the marks assigned for these various elements. For the most part there is agreement amongst five or six of them, and there are a few cases where seven, eight, nine, and ten examiners agree. The distribution of agreements in regard to the different elements is very much alike.

178. The following Table illustrates the disagreement amongst the examiners :—

TABLE 49

Number of Candidates

Number of different grades into which the candidates are placed by the several examiners	Vocabulary	Accuracy	Crafts-manship	Con-sistency	Com-pleteness	Substance	Quality
1	0	1	0	0	1	1	0
2	33	40	39	19	26	42	33
3	84	83	88	75	101	91	92
4	33	24	23	51	21	16	25
5	0	2	0	5	1	0	0
	150	150	150	150	150	150	150

179. For the most part the examiners agree in classifying the candidates in three different grades only ; thus they distribute their marks between, say, 1, 3 and 5, or between 0, 1 and 3, or between 3, 5 and 7. There are only a few cases where the whole gamut is employed by the examiners, but there are a number of cases of four grades being employed by the examiners.

While there is a distinct lack of agreement between the examiners, the marks are certainly *not* awarded in a purely haphazard fashion. For, if they were, we should expect to find many cases where the five marks, 0, 1, 3, 5, 7, were awarded by two examiners each out of the ten ; yet such a case never arises at all, and there are in fact very few cases where all five possible marks are awarded to any one candidate.

180. That the examiners have different standards is evident from Table 50 below, which shows to what number of candidates the different examiners have awarded the marks 0, 1, 3, 5, 7 in respect of the different elements :—

TABLE 50

Marks awarded for Vocabulary

	Examiner									
Marks	A	B	C	D	E	F	G	H	J	K
	Number of Candidates									
0	1								5	
1	30	8	6	3	3	2	17	5	37	2
3	74	63	47	34	11	37	71	59	67	89
5	36	75	52	73	49	69	54	66	31	45
7	9	4	45	40	87	42	8	20	10	14
Total	150	150	150	150	150	150	150	150	150	150

Marks awarded for Accuracy

	Examiner									
Marks	A	B	C	D	E	F	G	H	J	K
	Number of Candidates									
0	3	1	1			10	1	4	9	
1	5	13	2	11	9	9	12	15	37	20
3	42	76	32	42	17	32	24	33	59	72
5	86	59	95	77	62	68	94	82	36	56
7	14	1	20	20	62	31	19	16	9	2
Total	150	150	150	150	150	150	150	150	150	150

Marks awarded for Craftsmanship

	Examiner									
Marks	A	B	C	D	E	F	G	H	J	K
	Number of Candidates									
0	2	2			2			1	6	
1	17	13	9	2	16	12	25	6	30	11
3	97	83	51	47	32	57	60	77	64	92
5	30	50	63	73	62	68	61	55	40	42
7	4	2	27	28	38	13	4	11	10	5
Total	150	150	150	150	150	150	150	150	150	150

Marks awarded for Consistency

Marks	Examiner A	B	C	D	E	F	G	H	J	K
	Number of Candidates									
0								3	12	2
1		8	2	1	2	1	3	17	38	22
3	27	68	38	21	12	17	16	56	51	58
5	82	70	78	95	51	60	88	48	43	66
7	41	4	32	33	85	72	43	26	6	2
Total	150	150	150	150	150	150	150	150	150	150

Marks awarded for Completeness

Marks	Examiner A	B	C	D	E	F	G	H	J	K
	Number of Candidates									
0					2			4	5	
1	1	6		4	7	2	10	17	33	17
3	29	41	18	37	16	22	53	57	70	74
5	90	97	87	85	59	73	79	48	35	53
7	30	6	45	24	66	53	8	24	7	6
Total	150	150	150	150	150	150	150	150	150	150

Marks awarded for Substance

Marks	Examiner A	B	C	D	E	F	G	H	J	K
	Number of Candidates									
0									8	
1		5		6	1	9	5	14	35	3
3	25	51	38	53	31	46	57	37	59	87
5	62	83	84	72	72	66	71	68	38	59
7	63	11	28	19	46	29	17	31	10	1
Total	150	150	150	150	150	150	150	150	150	150

Marks awarded for Quality

Marks	Examiner A	B	C	D	E	F	G	H	J	K
	Number of Candidates									
0	1							1	9	
1	8	18	3	14	25	15	34	32	35	18
3	60	90	60	62	51	61	76	52	67	98
5	66	39	57	59	53	48	34	45	26	29
7	15	3	30	15	21	26	6	20	13	5
Total	150	150	150	150	150	150	150	150	150	150

181. A cursory glance at the above Table is sufficient to make one realise the difference between the standards of the different examiners. Thus, in marking for Vocabulary, Examiner E awards 87 full marks of 7, while Examiner B only awards 4; in marking for Accuracy, Examiner E gives full marks 62 times, while B only awards one candidate the full marks; in marking for Consistency, E gives full marks to 85 (more than half) of the candidates, while K gives full marks to only 2 candidates.

We find again that D never allots a zero, C and G each give 1 zero, K gives 2, B gives 3, and E gives 4 zeros, while J gives 54 zeros. Examiner K, who only gives 2 zeros, gives only 35 maximum marks out of the total number of 1,050; his marks are mostly 1, 3 and 5. Examiner B is also sparing of extreme marks; he only gives 3 zeros and 31 maximum marks.

182. The average marks awarded to the one hundred and fifty candidates indicate the differences of standard adopted by the examiners :—

TABLE 51

AVERAGE MARKS FOR ENGLISH PAPER, PART A

	Examiner									
	A	B	C	D	E	F	G	H	J	K
Vocabulary	3·30	4·00	4·81	5·00	5·93	5·01	3·71	4·35	3·09	3·95
Accuracy	4·39	3·62	4·75	4·41	5·36	4·41	4·58	4·24	3·05	3·53
Craftsmanship	3·24	3·51	4·44	4·69	4·59	4·09	3·59	3·93	3·28	3·55
Consistency	5·19	3·93	4·87	5·13	5·92	5·71	5·28	4·05	2·99	3·60
Completeness	4·99	4·37	5·36	4·72	5·41	5·36	4·13	3·97	3·11	3·64
Substance	5·51	4·33	4·87	4·39	5·17	4·53	4·33	4·55	3·15	3·77
Quality	4·15	3·36	4·52	4·00	3·93	4·13	3·16	3·69	3·05	3·28

RANGE OF AVERAGES FOR THE DIFFERENT ELEMENTS

	Highest	Lowest	Range
Vocabulary	5·93	3·09	2·84
Accuracy	5·36	3·05	2·31
Craftsmanship	4·69	3·24	1·45
Consistency	5·92	2·99	2·93
Completeness	5·41	3·11	2·30
Substance	5·51	3·15	2·36
Quality	4·52	3·05	1·47

183. For Vocabulary and Consistency the range of the average marks is nearly 3 marks (out of a maximum of 7); and in each case the extreme examiners are E and J. These wide differences between the average marks can only mean that these two examiners have fundamentally different things in mind when they are assessing these elements in the scripts.

184. In order to see whether the differences of standard between the different examiners vary with the element in composition

which they are estimating, or remain constant in judging the different elements, we have, in Table 52 below, placed the examiners in order according to the average marks which they have given to the whole one hundred and fifty candidates for each element. The examiners with the highest averages have been placed first in order.

TABLE 52

EXAMINERS IN THE ORDER OF THE AVERAGE MARKS ASSIGNED BY THEM TO THE DIFFERENT ELEMENTS

Examiner	Vocabulary	Accuracy	Craftsmanship	Consistency	Completeness	Substance	Quality
A	9	6	10	4	4	1	2
B	6	8	8	8	6	7½	7
C	4	2	3	6	2½	3	1
D	3	4½	1	5	5	6	4
E	1	1	2	1	1	2	5
F	2	4½	4	2	2½	5	3
G	8	3	6	3	7	7½	9
H	5	7	5	7	8	4	6
J	10	10	9	10	10	10	10
K	7	9	7	9	9	9	8

185. We see that several examiners keep a fairly constant position in the lists ; thus E and F are consistently high in their marking, and B, J, and K are consistently low. The diagram on p. 90 shows the averages of the different examiners in the form of a graph.

186. We have already pointed out (para. 181 above) how some examiners crowd their marks into the central groups, 1, 3 and 5, and avoid the extreme marks 0 and 7. It is convenient to measure the spread of the marks by calculating their mean deviations[1] shown in the following Table :—

TABLE 53

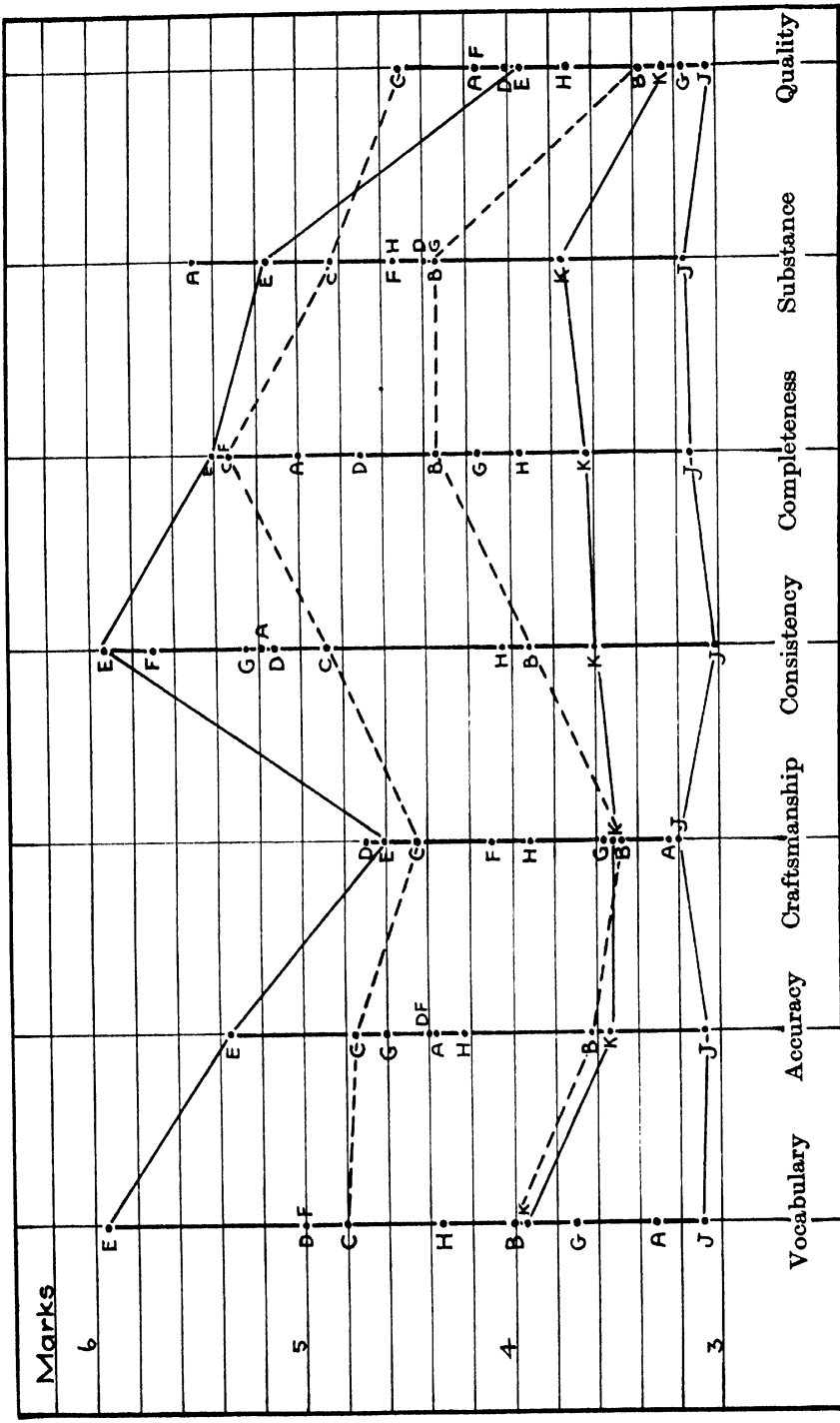
MEAN DEVIATIONS OF THE MARKS OF THE DIFFERENT EXAMINERS FOR THE DIFFERENT ELEMENTS

	Examiner									
	A	B	C	D	E	F	G	H	J	K
Vocabulary	1·26	1·16	1·44	1·07	1·24	1·11	1·28	1·28	1·31	1·20
Accuracy	1·18	1·13	0·91	1·29	1·36	1·60	1·14	1·42	1·41	1·19
Craftsmanship	0·90	1·09	1·39	1·16	1·56	1·33	1·33	1·24	1·41	1·05
Consistency	0·99	1·16	1·05	0·82	1·22	1·24	0·99	1·63	1·48	1·32
Completeness	0·82	1·02	0·98	1·05	1·40	1·16	1·22	1·63	1·24	1·23
Substance	1·25	1·13	0·94	1·25	1·12	1·37	1·24	1·42	1·45	1·01
Quality	1·31	1·00	1·36	1·39	1·61	1·55	1·14	1·67	1·36	0·91

[1] The mean deviation of a series of numbers is the average of their differences from their average.

SPECIAL PLACE EXAMINATION (I) : ENGLISH ESSAY

AVERAGES



187. As examples of marks with low mean deviations we may select those of Examiner A for the element "Completeness" He distributed his marks as follows :—

Marks	1	3	5	7
No. of Candidates	1	29	90	30

Mean Deviation = 0·82

Again, Examiner D distributed his marks for "Consistency" as follows :—

Marks	1	3	5	7
No. of Candidates	1	21	95	33

The Mean Deviation is again only 0·82

On the other hand, the distribution of Examiner H's marks for "Quality" is a much wider one :—

Marks	0	1	3	5	7
No. of Candidates	1	32	52	45	20

Mean Deviation = 1·67

and the distribution of Examiner F for "Accuracy" is again wide :—

Marks	0	1	3	5	7
No. of Candidates	10	9	32	68	31

Mean Deviation = 1·60

188. Some of the examiners are fairly consistent in their spreading of the marks; thus Examiner B's mean deviations range only from 1·00 to 1·16; Examiner J's mean deviations range only from 1·24 to 1·48. Examiner J consistently spreads his marks more than does B. We can better appreciate the differences between the examiners in this respect if we place them again in order, putting first those with the largest mean deviations :—

TABLE 54

EXAMINERS IN THE ORDER OF THE MEAN DEVIATIONS OF THEIR MARKS FOR THE DIFFERENT ELEMENTS

Examiner	Vocabulary	Accuracy	Crafts-manship	Consist-ency	Complete-ness	Substance	Quality
A	5	7	10	8½	10	4½	7
B	8	9	8	6	8	7	9
C	1	10	3	7	9	10	5½
D	10	5	7	10	7	4½	4
E	6	4	1	5	2	8	2
F	9	1	4½	4	6	3	3
G	3½	8	4½	8½	5	6	8
H	3½	2	6	1	1	2	1
J	2	3	2	2	3	1	5½
K	7	6	9	3	4	9	10

At one extreme are Examiners H and J, who spread their marks fairly widely, while at the other are B and A.

189. Examiners vary among themselves in their power of discriminating between the different candidates who are approximately at the same level of excellence. We may remind the reader that in a competitive examination, in which success depends on the results in more than one subject, the importance of a subject as a factor in the success of the candidate depends not only on the maximum assigned to it, but on the spreading of the marks in that subject as compared with the spreading in other subjects. To take an extreme instance: If all the candidates received between 50% and 55% of the marks in one subject, whereas in another subject, for which the maximum was the same, the marks varied from 10% to 80%, the marks in the second subject would be of far more importance in determining the place of a candidate than those in the first subject.

190. It is interesting, therefore, now to consider the difference in the spread of the marks, not from examiner to examiner, but from element to element. The following Table shows the averages of the mean deviations of the examiners for each element:—

Vocabulary	Accuracy	Crafts-manship	Consist-ency	Complete-ness	Substance	Quality
1·24	1·26	1·25	1·19	1·18	1·22	1·33

It would seem that examiners on the whole found it easier to discriminate between the candidates in respect of Quality than in respect of Consistency or Completeness.

191. We should safeguard ourselves by adding here that we are not at all convinced that the analysis of a composition into the seven elements adopted in the marking-scheme is the best possible, or that it can even be regarded as satisfactory. It was chosen as one that has been actually used, and we wished to investigate the kind of results that it yields. The determination of the mean deviation of the marks for each element is an important factor in judging of the scheme.

192. We have seen that the different examiners adopt different standards in marking for the seven elements of the composition separately. It was to be expected, therefore, that the total marks should show the same kind of difference of standard.

Thus in marking for the various elements, Examiner J's average is about 2 marks lower than E's on a maximum of 7. For the essays as a whole, with a maximum of 50, the difference of the average marks of the two examiners is 14 (seven times as great), the actual average of J being 22 and of E, 36.

193. We discuss in the next chapter the question whether the analysis of a composition into "elements" in this way yields more consistent results than marking by impression.

SECTION V. ENGLISH, PART B

194. Part B of the English paper contained four questions, mainly dealing with the sense of a passage, the sense of phrases, or the sense of single words. The maximum for Qn. 1 was 14 marks, and for each of the other questions, 12 marks.

195. We quote from the paper Qns. 1 and 4. The general directions given with regard to the paper as a whole, and the specific directions given with regard to Qns. 1 and 4 are given in Appendix I to this chapter (pp. 112, 113).

Qn. 1 reads as follows :—

Read the following :—

Men till the fields at Littleport,
The spreading fields and low
And as they toil amid the soil
I wonder if they know
That where they drop the yellow grain
An ocean used to flow,
And little ships to little quays
Came gladly after tossing seas,
And sailors laughed and took their ease
Long, long ago.

Answer the following :—

(*a*) What is the chief work of the men of Littleport now ?
(*b*) Express in your own words "drop the yellow grain," "quays," "tossing seas."
(*c*) Could you sail in a ship to Littleport now ? Why, or why not ?
(*d*) *Quote* another phrase for "till the fields."
(*e*) Describe in a few words what Littleport was like "long, long ago."

Qn. 2 is a request to write sentences showing the meanings of four simple phrases.

Qn. 3 is a request to write sentences showing the uses of three pairs of words having some similarity of form.

Qn. 4 reads as follows :—

Describe, each in one word, the following :—

(*a*) A man who went to the ends of the earth ;
(*b*) A woman who looks after a Post Office ;
(*c*) A man who cultivates flowers ;
(*d*) A boy who sells the *Bugle*[1] and the *Evening Mail*[1] ;
(*e*) A girl who looks after tiny children ;
(*f*) A man who makes wooden things.

196. It is true that a certain latitude was given to examiners in regard to fresh points which might arise after the marking-

[1] These names are fictitious.

scheme had been settled ; but the discussion with the examiners on the basis of 100 trial papers (see para. 140 above and footnote) had reduced the number of " fresh points " on so simple a paper to a small one, and it was anticipated that the agreement between the examiners would be close. But it will be seen that there was little agreement except in regard to Qn. 4.

197. The number of possible marks (including zero) is 15 for Qn. 1, and 13 for each of the other questions ; and the different marks given by the examiners are so numerous that a table showing the number of times each combination occurred would be unwieldy. But Table 55 below gives a conspectus of the agreement of the different examiners.

TABLE 55

NUMBER OF CASES OF AGREEMENT BETWEEN THE EXAMINERS

	Questions			
	1	2	3	4
	Number of Candidates			
10 examiners agree	1	0	0	66
9 ,, ,,	3	0	2	22
8 ,, ,,	6	1	3	11
7 ,, ,,	7	5	3	9
6 ,, ,,	20	5	6	17
5 ,, ,,	41	7	10	13
4 ,, ,,	38	26	34	8
All the remainder, where there is never more agreement than between two or three examiners	34	106	92	4
	150	150	150	150

198. It will be seen that of the answers to Qn. 4, sixty-six receive the same mark from all the ten examiners.

The marks of the answers to Qns. 2 and 3 show the least agreement ; in no case do all ten examiners agree, and in the case of Qn. 2 in no case do nine examiners agree.

199. In the answers to Qn. 1 there are 78 cases in which five or more examiners agree. But in the answers to Qn. 2 there are only 18 cases in which five or more agree.

Thus it is only in dealing with the answers to Qn. 4 that there is anything approaching a general agreement.

200. Let us now turn to the ranges of the marks awarded by the different examiners to the same answer. The distribution of the ranges of marks is shown in Table 56 below.

TABLE 56

DISTRIBUTION OF RANGE OF MARKS

(ENGLISH, PART B)

		Question			
		1	2	3	4
Range	Max.	14	12	12	12
0		1	0	0	66
1		9	3	2	32
2		37	5	9	36
3		35	7	9	13
4		33	16	22	3
5		15	16	21	
6		13	30	18	
7		3	26	20	
8		4	30	14	
9			7	18	
10			6	8	
11			2	6	
12			2	3	
	Total	150	150	150	150
Average Range		3·51	6·33	6·29	1·03

201. In a few cases where the maximum is 12, the range is also 12, i.e. the marks given by different examiners vary from 0 to the maximum. The details of the cases are as follows :—

Qn. 2. Candidate No. 41 received the following marks from Examiners A to K respectively 5, 9, 6, 8, 10, 8, 0, 4, 12, 7

Candidate No. 102 received the following marks 2, 7, 7, 12, 9, 2, 0, 5, 2, 4

Qn. 3. Candidate No. 89 ,, ,, ,, ,, 5, 9, 2, 12, 7, 0, 0, 6, 6, 5

Candidate No. 116 ,, ,, ,, ,, 4, 5, 0, 12, 8, 0, 0, 5, 6, 6

Candidate No. 148 ,, ,, ,, ,, 8, 6, 3, 12, 10, 0, 3, 9, 7, 7

202. In the case of Qn. 4, where there is most agreement, the average range is only 1 mark ; for Qn. 1 it is 3½ ; and for Qns. 2 and 3 it is 6 marks, that is, half the maximum.

There is thus no semblance of agreement in the marking of the answers to Qns. 2 and 3.

203. We turn now to the distribution of the marks for the whole set of scripts, taking each question separately, as set out in Table 57 below :—

TABLE 57

DISTRIBUTION OF MARKS

(ENGLISH, PART B)

Question 1

	Examiner									
Marks	A	B	C	D	E	F	G	H	J	K
0	1	3	1	1	1	1	5	3	1	3
1	4	4	3	5	1	5	6	3	1	2
2	2	3	4	7	4	4	4	8	7	5
3	8	1	4	5	3	5	5	3	3	8
4	8	7	8	7	2	3	8	7	3	4
5	4	1	5	8	3	5	10	9	3	14
6	12	11	9	12	6	7	11	11	9	9
7	12	14	14	20	10	16	21	17	12	18
8	23	18	22	14	8	21	17	15	13	18
9	25	27	22	21	18	22	22	23	17	15
10	17	22	24	16	34	17	20	20	26	25
11	20	18	14	13	27	26	11	22	33	17
12	10	16	14	13	22	12	7	5	10	7
13	3	4	5	6	10	5	3	3	8	4
14	1	1	1	2	1	1	–	1	4	1
Total	150	150	150	150	150	150	150	150	150	150

Question 2

	Examiner									
Marks	A	B	C	D	E	F	G	H	J	K
0				1			2	1	1	
1	2			1		1			1	2
2	2					6	5	1	1	
3	9			1		6	8	5	2	3
4	21	2	4	7	1	15	12	11	12	27
5	22	5	4	3	2	4	11	14	14	31
6	24	4	6	3	1	25	18	16	25	26
7	25	14	7	6	3	19	17	21	33	21
8	18	21	24	5	12	22	19	24	26	20
9	9	52	20	15	17	22	16	27	11	8
10	12	20	27	10	18	12	13	12	7	4
11	5	11	20	5	55	8	15	10	13	6
12	1	21	38	93	41	10	14	8	4	2
Total	150	150	150	150	150	150	150	150	150	150

TABLE 57—*continued*

Question 3

Marks	Examiner A	B	C	D	E	F	G	H	J	K
0			2			9	7			
1			1			11	5			
2		1	5			16	10	1		1
3	4	1	9			21	21	1	1	5
4	15	4	11		1	11	17	4	5	8
5	19	6	9			16	15	3	7	19
6	27	25	18	8	22	20	30	13	27	21
7	33	24	16	6	2	8	15	15	21	25
8	22	26	24	17	58	14	13	26	34	23
9	9	23	17	15	2	11	1	25	22	26
10	17	15	17	32	35	9	15	28	24	16
11	3	14	8	17		1	1	18	5	2
12	1	11	13	55	30	3		16	4	4
Total	150	150	150	150	150	150	150	150	150	150

Question 4

Marks	Examiner A	B	C	D	E	F	G	H	J	K
0	1	1	1	1	1	1	1	1	1	1
1										
2	2	1	2	2		1	2	2	1	2
3						1				
4	2	2	4	2	2	1	2	2	2	3
5	4	2	2	3	2	3	4	3	3	2
6	13	8	11	7	8	8	11	9	9	14
7	5	7	4	6	3	8	7	9	7	3
8	33	28	34	25	25	29	32	28	28	30
9	11	8	7	14	15	15	12	12	13	14
10	41	45	49	44	44	41	42	40	46	44
11	5	13	6	11	12	6	5	13	4	7
12	33	35	30	35	38	36	32	31	36	30
Total	150	150	150	150	150	150	150	150	150	150

204. *Qn.* 1. There are no striking differences between the different examiners in the distribution of the marks for this question. All the possible marks from 0 to 14 are used.

Qn. 2. Here there are very great differences in the distribution of the marks. Examiners B, C and E never use any mark below 4; Examiners A, J and K are very sparing in the award of the maximum mark (12); whereas D awards the maximum to ninety-three candidates, and C and E award it to thirty-eight and forty-one candidates respectively.

Qn. 3. We find similar peculiarities in the distribution for this question. Examiner E nearly always uses 6, 8, 10 or 12 marks. Examiner D awards the maximum to fifty-five candidates, while A, F, G, J and K give the maximum to four candidates or fewer. Examiner D does not allot any mark below

6, while F gives a mark below 6 to eighty-four candidates (more than half the total number).

It is obvious that there are between the examiners fundamental differences in their interpretation, either of the scripts or of the instructions, in dealing with Qns. 2 and 3.

Qn. 4. As with Qn. 1, we have here distributions of the marks which are alike for all the examiners.

205. We now turn to the differences of standard of the examiners as indicated by differences of their averages, and to the differences of their distribution as indicated by the mean deviations of their marks.

206. TABLE 58

AVERAGE MARKS FOR QUESTIONS IN ENGLISH, PART B, AND ORDER OF EXAMINERS ARRANGED ACCORDING TO THE MAGNITUDE OF THEIR AVERAGES

Question	Averages				Order of Examiners according to the magnitude of the averages Question			
	1 (Max. 14)	2 (Max. 12)	3 (Max. 12)	4 (Max. 12)	1	2	3	4
Examiner								
A	8·08	6·39	6·88	9·13	6	9	8	9½
B	8·49	9·09	8·13	9·51	3	4	4	2
C	8·31	9·59	7·37	9·17	5	3	6	7
D	7·87	10·35	10·19	9·45	7	2	1	3
E	9·45	10·45	8·95	9·67	1	1	2	1
F	8·43	7·31	5·03	9·35	4	7	10	5
G	7·35	7·49	5·27	9·13	10	6	9	9½
H	7·79	7·57	8·81	9·27	8½	5	3	6
J	8·99	7·17	7·87	9·37	2	8	5	4
K	7·79	6·26	7·29	9·16	8½	10	7	8

The following Table shows the highest and lowest averages for each question :—

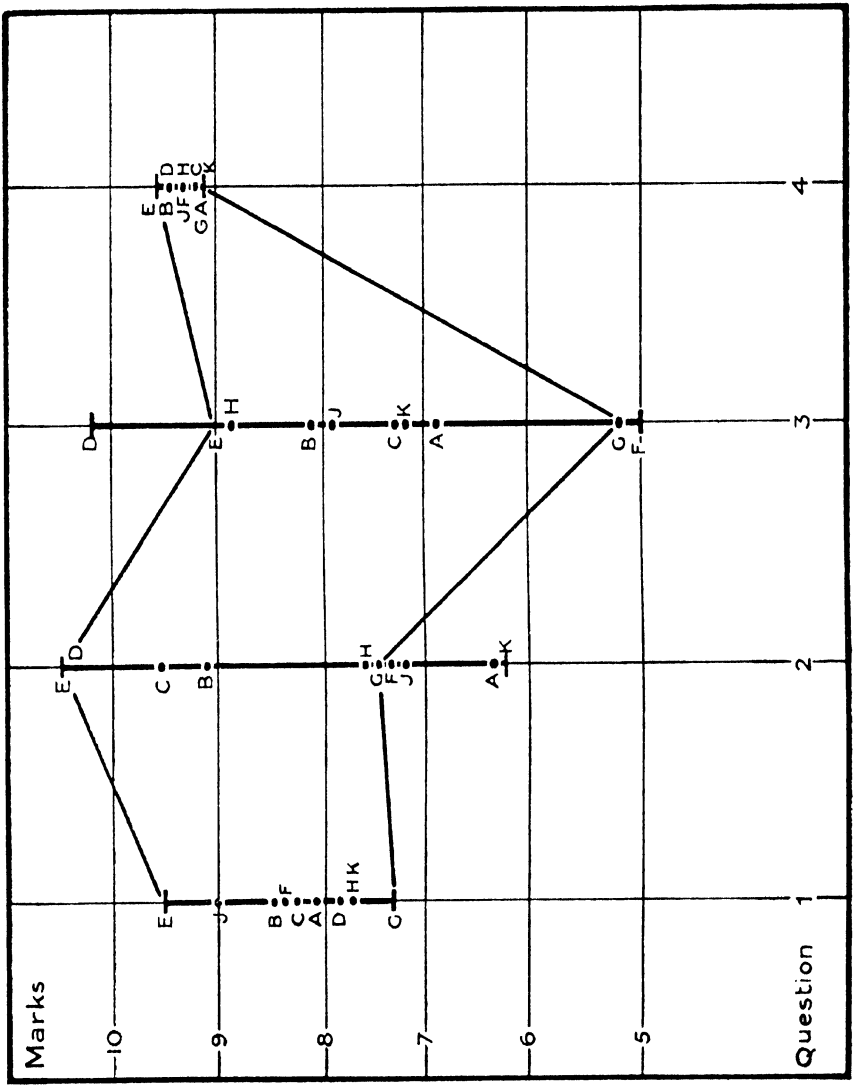
TABLE 59

Question	Highest	Lowest	Range
1. Max. = 14	9·45	7·35	2·10
2. Max. = 12	10·45	6·26	4·19
3. Max. = 12	10·19	5·03	5·16
4. Max. = 12	9·67	9·13	0·54

207. The wide variations of the averages compared to the maximum for each question emphasise again the differences between the marks of the different examiners. The second part of Table 58 above shows that there is some degree of consistency of standard in the marking of the examiners for the different questions, since Examiner E is obviously higher in his marking, almost throughout, than the majority of his colleagues, and Examiner G is on the whole lower (though not so markedly in regard to Qn. 2). On the other hand, the position of Examiner J in the Table varies considerably.

208. The following graph shows the averages in a striking form.

AVERAGES



209. It is interesting to compare the averages of the marks allotted by the different examiners for Part A (Essay) with those for Part B, and with the averages for Parts A and B combined. Table 60 below gives these averages and also shows the order of the Examiners arranged according to their averages:—

TABLE 60

1	2	3	4	5	6	7
Examiner	English, Part A Max. 50	English, Part B Max. 50	Order of Examiners according to averages for Part A	Order of Examiners according to averages for Part B	English, Parts A & B Averages for Parts A & B combined	Order of Examiners according to averages for Parts A & B combined
A	30·77	30·48	5	8	61·25	7
B	27·12	35·22	8	3	62·34	5
C	33·62	34·45	2	4	68·07	3
D	32·34	37·86	4	2	70·20	2
E	36·31	38·52	1	1	74·83	1
F	33·24	30·12	3	9	63·36	4
G	28·78	29·24	6½	10	58·02	8
H	28·78	33·44	6½	5	62·22	6
J	21·72	33·40	10	6	54·04	10
K	25·32	30·50	9	7	55·12	9

210. It will be seen that Examiner E consistently gives high marks throughout. J and K give low marks for the Essay. The difference between the average mark of E for the whole paper (74·8) and that of J (55·1) is 20 out of a maximum of 100. Of this, 15 comes from the Essay and the other 5 from Part B.

211. We now come to the mean deviations of the examiners in marking English, Part B, shown in Table 61 below, together with the order of the examiners, arranged according to the magnitude of their mean deviations:—

TABLE 61

ENGLISH—PART B

	Mean Deviations				Order of Examiners according to magnitude of Mean Deviation			
	Question				Question			
	1	2	3	4	1	2	3	4
Examiner	(Max. 14)	(Max. 12)	(Max. 12)	(Max. 12)				
A	2·32	1·87	1·58	1·87	8	5	8	1
B	2·30	1·33	1·74	1·72	9	9	4½	9
C	2·33	1·74	2·35	1·82	6½	6½	2	3
D	2·63	2·09	1·51	1·74	1	3	9	7½
E	2·09	1·25	1·71	1·59	10	10	6½	10
F	2·37	2·15	2·56	1·76	5	2	1	6
G	2·54	2·43	2·19	1·83	2	1	3	2
H	2·52	1·97	1·71	1·81	3	4	6½	4
J	2·33	1·74	1·50	1·74	6½	6½	10	7½
K	2·51	1·70	1·74	1·80	4	8	4½	5

212. Some of the examiners show consistency in their spreading of the marks, e.g. E, whose mean deviation is low throughout, and G, whose mean deviation is high throughout. Other examiners, e.g. Examiners C and D, vary from question to question.

SECTION VI. ENGLISH, PART B

DETAILED EXAMINATION OF THE MARKS AWARDED FOR PARTS OF A QUESTION

213. Up to the present we have discussed only the marking of questions as a whole and the agreement and differences of the marks of different examiners. We shall get to closer grips with the problem of English, Part B, by comparing the marks of the examiners for the different parts of a question, and we select for this Qn. 1. Qn. 1 of Part B requires five answers, (*a*), (*b*), (*c*), (*d*), (*e*), each of which has a maximum of 3 marks, except (*d*), which carries only 2 (see para. 195 above).

214. Table 62 below shows the number of cases in which examiners agree :—

TABLE 62

AGREEMENTS BETWEEN EXAMINERS IN MARKING THE DIFFERENT PARTS OF QN. 1, ENGLISH, PART B

Examiners	Part of Qn. 1 (*a*)	(*b*)	(*c*)	(*d*)	(*e*)
	No. of cases				
10 agree	64	12	42	88	31
9 ,, the rest not agreeing	39	22	27	47	29
8 ,, ,, ,,	14	34	12	7	18
7 ,, ,, ,,	5	24	17	4	26
6 ,, ,, ,,	9	22	18	1	12
5 ,, ,, ,,	7	20	16		13
	138	134	132	147	129
5, 5	1	5	1		2
4 at most agree	5	10	8		13
3 ,, ,,					2
	144	149	141	147	146

The upper part of the above Table shows the number of scripts in regard to which agreement is reached by five examiners, the rest differing among themselves as well as from the first five.

In the lower part of the Table, the figures "5, 5" mean that five agree on a certain mark and the other five on another mark. The other cases shown in the Table refer to scripts as to which three or at most four examiners agree in assigning the same mark ;

it includes one case where three examiners award one mark, three others award another mark, two others award a third mark, and the remaining two award a fourth mark.

215. Part (*d*) of the question ("*Quote another phrase for 'till the fields'*") yields most agreement, but it is to be remembered that there are only three possible marks (0, 1, 2); even on this question we find that Examiner C gave 3 marks, i.e. one more than the maximum, to six candidates; and D gave 3 marks to three candidates.[1]

The answers to this part are on the whole poor; a majority of the examiners award 0 marks to one hundred and five candidates.

216. Of the other parts, Qn. 1 (*a*) ("*What is the chief work of the men of Littleport now?*") and Qn. 1 (*c*) ("*Could you sail in a ship to Littleport now? Why or why not?*") yield most agreement among the examiners; and Qn. 1 (*b*) ("*Express, in your own words, 'drop the golden grain', 'quays', 'tossing seas'*") and Qn. 1 (*e*) ("*Describe in a few words what Littleport was like 'long, long ago'*") the least. As to Qn. 1 (*b*) there is complete agreement in only 12 cases out of 149.

217. It is clear that the different parts of the question contribute unequally towards the total amount of discrepancy; part (*a*) is easier to mark than part (*b*).

Even with such detailed instructions as were given, an elementary question of this kind is far from being "fool-proof."

218. The different parts of Qn. 1 (see para. 195) would seem at first sight to offer little opportunity for differences of opinion among the examiners, yet the facts show that there is a great deal.

It seemed worth while to investigate whether there is any kind of statistical regularity in these differences. In what follows we study the number of instances in which an examiner differs from the majority of his colleagues. In this study we shall leave out the cases in which five examiners agree on one mark and the other five on another mark.

219. *Qn.* 1 (*a*).—Table 62 shows that there are 74 (= 138 − 64) cases in which one or more examiners differ from a majority. The following Table gives particulars of the dissentient[2] examiners and of the number of candidates involved.

[1] If a detailed mark-sheet, such as we used, is employed, errors of this kind on the part of examiners should be detected; with a less detailed mark-sheet they might easily escape attention.

[2] Although for convenience we have used the word "dissentient" to denote an examiner whose marks disagree with those of the majority, it is to be remembered that the examiners had no knowledge whatever of each other's marks.

TABLE 63

Qn. 1 (*a*)

Number of cases in which individual examiners disagree with the majority

No. of Examiners who agree	No. of cases of disagreement										Total No. of cases of disagreement	No. of candidates concerned
	A	B	C	D	E	F	G	H	J	K		
9		3	2	2	3	1		1	3	24	39	39
8			1	9	2		2	3	1	10	28	14
7	1	2	1	2	1	1		2	1	4	15	5
6	6	4	2	5	4	2	2	4	4	3	36	9
5	4	4	2	2	4	4	4	3	3	5	35	7
	11	13	8	20	14	8	8	13	12	46	153	74

220. Thus, out of the 39 cases in which nine examiners agree and there is a single dissentient, the dissentient is Examiner K in 24 cases.

Again, out of the 14 cases in which eight examiners agree and there are two dissentients, K is a dissentient in 10 cases. On the other hand, C, F and G are most often on the side of the majority.

221. The next point to consider is whether a dissentient examiner tends to mark higher or lower than the majority, or whether he marks sometimes high, sometimes low. Table 64 shows the number of cases of disagreement of the several examiners with the majority, and the extent of the disagreement above or below the mark of the majority :—

TABLE 64

Qn. 1 (*a*)

Extent and sign of disagreement in cases in which individual examiners disagree with the majority

Extent and sign of disagreement expressed in marks	Number of cases of disagreement									
	A	B	C	D	E	F	G	H	J	K
+3					1					
+2		2			2					
+1	7	6	1	1	10	1	2	1	5	
−1	3	5	7	17	1	7	2	10	6	44
−2	1			2				2		2
−3							4		1	
+	7	8	1	1	13	1	2	1	5	
−	4	5	7	19	1	7	6	12	7	46
Total	11	13	8	20	14	8	8	13	12	46

222. From this we see that Examiner K, who disagrees most often with the majority, consistently gives a lower mark. In all cases but one, when they disagree with the majority, C, D, F and H give a lower mark and E gives a higher mark than the majority ; while A, B and J mark sometimes high and sometimes low.

223. *Qn.* 1 (*b*).—We give in Tables 65 and 66 below a similar analysis with regard to Qn. 1 (*b*) :—

TABLE 65

Qn. 1 (*b*)

Number of cases in which individual examiners disagree with the majority

Number of Examiners who agree	Examiner										Total No. of cases of disagreement	No. of candidates concerned
	A	B	C	D	E	F	G	H	J	K		
9		4		4	1	2	10		1		22	22
8	1	13	1	16	5	2	18	7	4	1	68	34
7	4	9	6	11	8	6	13	9	4	2	72	24
6	9	5	6	7	14	10	7	8	12	10	88	22
5	8	3	8	17	14	10	14	11	9	6	100	20
	22	34	21	55	42	30	62	35	30	19	350	122

In this part of the question, K, who dissented most often for part (*a*), dissents least often, and D and G dissent most often.

224.

TABLE 66

Qn. 1 (*b*)

Extent and sign of disagreement in cases in which individual examiners disagree with the majority

Extent and sign of disagreement expressed in marks	Examiner									
	A	B	C	D	E	F	G	H	J	K
+2		1		4	2					
+1	8	6	15	25	38	23	1	1	22	5
−1	14	27	6	20	2	7	56	31	7	14
−2				6			5	3	1	
+	8	7	15	29	40	23	1	1	22	5
−	14	27	6	26	2	7	61	34	8	14
Total	22	34	21	55	42	30	62	35	30	19

G and H, in all cases but one of their dissent, mark lower than the majority. E nearly always marks higher than the majority. D disagrees with the majority in many cases, and marks sometimes low, sometimes high, with almost equal frequency. A, B and K more often give a low mark, while C, F and J more often give a high mark.

225. TABLE 67

Qn. 1 (*c*)

Number of cases in which individual examiners disagree with the majority

Number of Examiners who agree	Examiner										Total No. of cases of disagreement	No. of candidates concerned
	A	B	C	D	E	F	G	H	J	K		
9	1	1	2	17			5		1		27	27
8	2	3	1	9	1	1	3	2	2		24	12
7	4	6	11	12	1	4	3	1	1	8	51	17
6	14	4	8	11	9	7	4	1	6	8	72	18
5	9	9	9	9	9	5	7	4	7	12	80	16
	30	23	31	58	20	17	22	8	17	28	254	90

Here again Examiner D disagrees most often, almost twice as often as any other examiner. H disagrees in only eight cases.

226. TABLE 68

Qn. 1 (*c*)

Extent and sign of disagreement in cases in which individual examiners disagree with the majority

Extent and sign of disagreement expressed in marks	Examiner									
	A	B	C	D	E	F	G	H	J	K
+3									1	
+2	3	1			6	2			3	
+1	11	3	5	3	13	4	2	3	6	4
−1	14	16	22	32	1	9	17	5	7	18
−2	2	2	3	21		1	3			6
−3		1	1	2		1				
+	14	4	5	3	19	6	2	3	10	4
−	16	19	26	55	1	11	20	5	7	24
Total	30	23	31	58	20	17	22	8	17	28

Examiner D almost always marks lower, and B, C, G and K generally mark lower, than the majority. E marks higher in every case but one. A, F, H and J vary in their manner of disagreeing.

227. TABLE 69

Qn. 1 (*d*)

Number of cases in which individual examiners disagree with the majority

Number of Examiners who agree	Examiner A	B	C	D	E	F	G	H	J	K	Total No. of cases of disagreement	No. of candidates concerned
9	1	7	3	27					9		47	47
8	2	1	3	4	1		1		2		14	7
7		4		4	2			1	1		12	4
6		1		1	1				1		4	1
	3	13	6	36	4		1	1	13		77	59

Here again Examiner D is the chief dissentient, while F and K are always with the majority.

228. TABLE 70

Qn. 1 (*d*)

Extent and sign of disagreement in cases in which individual examiners disagree with the majority

Extent and sign of disagreement expressed in marks	Examiner A	B	C	D	E	F	G	H	J	K
+2				2					2	
+1		12	6	34	4			1	9	
−1	3	1					1		2	
+		12	6	36	4			1	11	
−	3	1					1		2	
Total	3	13	6	36	4		1	1	13	

Here Examiner D consistently gives higher marks than the rest; B and J in most of the cases award higher marks.

229. TABLE 71

Qn. 1 (*e*)

Number of cases in which individual examiners disagree with the majority

Number of Examiners who agree	Examiner A	B	C	D	E	F	G	H	J	K	Total No. of cases of disagreement	No. of candidates concerned
9	2	7	2	11	4		1		2		29	29
8	7	6	2	7	6	2	3		2	1	36	18
7	12	9	5	8	8	6	10	7	9	4	78	26
6	5	4	5	5	7	6	3	3	3	7	48	12
5	5	6	9	5	8	7	8	8	7	2	65	13
	31	32	23	36	33	21	25	18	23	14	256	98

There is no examiner pre-eminent in disagreement, though A, B, D and E disagree more often than the rest.

230. TABLE 72

Qn. 1 (*e*)

Extent and sign of disagreement in cases in which individual examiners disagree with the majority

Extent and sign of disagreement expressed in marks.	Examiner									
	A	B	C	D	E	F	G	H	J	K
+3					1					
+2		5		1	5	1			3	
+1	2	20	10	4	23	4		3	16	7
−1	23	6	13	31	4	15	19	11	4	7
−2	6	1				1	5	4		
−3							1			
+	2	25	10	5	29	5		3	19	7
−	29	7	13	31	4	16	25	15	4	7
Total	31	32	23	36	33	21	25	18	23	14

Here Examiner G in his disagreement marks consistently lower than his colleagues ; and A, D, F and H mainly mark low. B, E and J mainly mark high ; and C and K sometimes mark high, sometimes low.

231. Let us now from these scattered Tables try to compare the characteristics of the different examiners.

In marking Qn. 1 (*a*) there were 153 occasions on which different examiners disagreed with the majority, and to these Examiner A contributed 11, i.e. 7·2%. We show in Table 73 below this and similar percentages for all the examiners for each part of Qn. 1.

TABLE 73

Qn. 1

PERCENTAGES OF TOTAL NUMBER OF DIFFERENCES FROM THE MAJORITY DUE TO EACH EXAMINER

Parts of Qn. 1	Examiner									
	A	B	C	D	E	F	G	H	J	K
(*a*)	7·2	8·5	5·2	13·1	9·2	5·2	5·2	8·5	7·8	30·1
(*b*)	6·3	9·7	6·0	15·7	12·0	8·6	17·7	10·0	8·6	5·4
(*c*)	11·8	9·1	12·2	22·8	7·9	6·7	8·7	3·1	6·7	11·0
(*d*)	3·9	16·9	7·8	46·8	5·2		1·3	1·3	16·9	
(*e*)	12·1	12·5	9·0	14·1	12·9	8·2	9·8	7·0	9·0	5·5
Average	8·3	11·3	8·0	22·5	9·4	5·7	8·5	6·0	9·8	10·4

D is thus the examiner who differs most frequently from the majority. K differs most in regard to part (*a*), but agrees well with the majority on the other parts of the question.

232. An examination of the records of the different examiners showing the number of occasions in which they respectively marked higher and lower than the majority is interesting.

The cases in which they marked higher and lower are denoted by " plus " and " minus " respectively.

233. Examiner A

Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	7	8	14	0	2
Minus	4	14	16	3	29
Total	11	22	30	3	31

Examiner A is inconsistent in his manner of disagreeing with the majority ; he sometimes marks high, sometimes low, except in part (*e*), when he almost consistently gives lower marks.

234. Examiner B

Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	8	7	4	12	25
Minus	5	27	19	1	7
Total	13	34	23	13	32

Examiner B disagrees in a fairly consistent manner ; parts (*b*) and (*c*) he mostly marks below the majority ; parts (*d*) and (*e*) higher than the majority ; in part (*a*) he is inconsistent.

235. Examiner C

Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	1	15	5	6	10
Minus	7	6	26	0	13
Total	8	21	31	6	23

In dealing with part (*e*), C marks high and low with almost equal frequency. He is fairly consistent in the other parts, marking low for parts (*a*) and (*c*), and high for (*b*) and (*d*).

236.

Examiner D					
Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	1	29	3	36	5
Minus	19	26	55	0	31
Total	20	55	58	36	36

Examiner D is very consistent in marking except in respect of part (*b*). He gives low marks for (*a*), (*c*) and (*e*), and high marks for (*d*).

237.

Examiner E					
Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	13	40	19	4	29
Minus	1	2	1	0	4
Total	14	42	20	4	33

Examiner E is most consistent. He almost always marks higher than the majority when he disagrees with them.

238.

Examiner F					
Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	1	23	6		5
Minus	7	7	11		16
Total	8	30	17		21

Examiner F is only fairly consistent in his marking of these parts. In part (*a*) he is consistently low; in part (*b*) he gives about three times as many high marks as low marks; in part (*e*) he reverses the process; and in part (*c*) he gives twice as many low as high marks.

239.

Examiner G					
Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	2	1	2	0	0
Minus	6	61	20	1	25
Total	8	62	22	1	25

Examiner G gives lower marks than the majority, when he disagrees with them, almost consistently.

240.

Examiner H					
Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	1	1	3	1	3
Minus	12	34	5	0	15
Total	13	35	8	1	18

Examiner H is consistent when marking parts (*a*) and (*b*) and fairly consistent when dealing with (*e*), but in the few cases of disagreement on (*c*) marks sometimes high, sometimes low.

241.

Examiner J					
Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	5	22	10	11	19
Minus	7	8	7	2	4
Total	12	30	17	13	23

Examiner J tends on the whole to mark high, but distributes his marks between high and low in a good many cases.

242.

Examiner K					
Part of question	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)
Plus	0	5	4		7
Minus	46	14	24		7
Total	46	19	28		14

In marking part (*e*) Examiner K distributes his marks equally between high and low ; in marking part (*a*) he is consistently low ; and in parts (*b*) and (*c*) he is generally low.

243. It will be seen that despite the attempts to secure uniformity in the marking of a comparatively simple question, the peculiar differences of examiners from the majority become apparent in their marking, and show themselves in three different ways :—

(i) by always or nearly always giving lower marks ;
(ii) by always or nearly always giving higher marks ;
(iii) by giving sometimes higher and sometimes lower marks.

We find from the preceding Tables that in dealing with different parts of Qn. 1, the several examiners act in the following ways.

Examiner	A	acts	according	to (i) and (iii)
,,	B	,,	,,	,, (i), (ii) and (iii)
,,	C	,,	,,	,, (i), (ii) and (iii)
,,	D	,,	,,	,, (i), (ii) and (iii)
,,	E	,,	,,	,, (ii)
,,	F	,,	,,	,, (i), (ii) and (iii)
,,	G	,,	,,	,, (i)
,,	H	,,	,,	,, (i) and (iii)
,,	J	,,	,,	,, (ii) and (iii)
,,	K	,,	,,	,, (i) and (iii)

We may, at this stage, refer to a previous discussion of Qn. 1 where it emerged that Examiner E had the highest average mark, and G had the lowest average, and D had the largest mean deviation (see paras. 206, 207 and 211 above).

244. The reasons for the differences are now apparent. When E differs from his colleagues, he awards higher marks; G in the same circumstances awards lower marks. Examiner D, we observed, was the greatest disturber of agreement, and this means, since his method of differing is sometimes to mark up, sometimes to mark down, that his spreading of the marks is greater than that of any of the other examiners.

245. The problem of spreading the marks in dealing with a paper like Part B, in which there should be very little latitude for different judgments, is entirely different from the corresponding problem in judging an essay like that set in Part A. Without a complete analysis answer by answer it is difficult to say whether a difference from the majority is due to closer attention to the detailed points than the majority have given, or to random marking.

It is generally recognised that not only the average mark but also the distribution should be analysed for each separate examiner when a large series of scripts is divided up among a number of assistant-examiners (see para. 148 above).

APPENDIX I TO CHAPTER VI

SPECIAL PLACE EXAMINATION (I)

ENGLISH PAPER, PART A

General Instructions for Examiners (Part A only)

1. The maximum number of marks is 50.

2. Seven marks are to be allotted under each of the headings set out on the mark-sheet, together with one additional mark for papers of outstanding merit.

3. Under each heading only the following marks are to be given: 0, 1, 3, 5, 7.

4. The following directions will explain in what way marks are to be allotted or deducted from the maximum under each of the headings specified.

(a) *Vocabulary.*—Mark for quantity and quality.

(b) *Accuracy.*—Deduct marks for technical errors in spelling, grammar, and punctuation. Capitals, full stops, and quotation marks should be specially considered.

(c) *Craftsmanship.*—Mark for skill in the use of phrases, the construction of sentences, and the formation of paragraphs.

(d) *Consistency* or "staying power."—Deduct marks for irrelevancies and confusions.

(e) *Completeness.*—Mark for unity of plan, with suitable beginning and ending. Marks for the title of Subject 1 should be given under this head.

(f) *Substance.*—Mark for number of ideas.

(g) *Quality.*—Mark for quality of ideas.

5. No deductions are to be made for lack of close attention to the questions set; these latter are merely designed to stimulate the child to write passages of connected prose.

6. In deducting marks take the length of the essay into consideration. Assume 200 words to be the average, and mark on the basis of the number of errors per 200 words.

7. If any fresh point of difficulty arises, each examiner must use his or her own discretion.

ENGLISH PAPER, PART B

General Instructions for Examiners (Part B only)

(1) Note that Direction 5 for Part A does not apply to Part B.

(2) Reasonable synonyms may be adopted throughout. If in doubt, give half marks.

(3) Allow no bonus for a specially good answer ; that is, the maximum mark for each question or each part of a question must not be exceeded.

(4) No deductions are to be made for faulty spelling in Part B.

(5) If any fresh point of difficulty arises, each examiner must use his or her own discretion.

Instructions regarding the several questions of Part B

Qn. 1. Maximum 14.

(*a*) Farming (3), agriculture (3), cultivating the land (3), tilling the soil (3), tilling (2), gardening (1).

(*b*) Sow (1), sowing (1), planting corn (1) ; wharves (1), jetties (1), piers (1), landing stages (1), docks (0), harbours (0) ; stormy weather (1), rough water (1).

(*c*) No, because it is dry land (3).

(*d*) Toil amid the soil (2), or drop the yellow grain (1).

For a quotation *not* in the given passage, e.g. " Plough the fields," from the well-known hymn, give (1) mark.

(*e*) A small seaport (3). If " small " is omitted give (2) marks. A port (2), a seaside town (1).

Qn. 4. Maximum 12 (2 each).

(*a*) Traveller (2), explorer (2), circumnavigator (1), discoverer (1), pioneer (0), Amundsen (0).

(*b*) Post Mistress (2), postwoman (1), superintendent (1), clerk (0).

(*c*) Gardener (2), florist (2), nurseryman (2), flower man (0), cultivator (0).

(*d*) Newsboy (2), newspaper boy (2), paper boy (2), hawker (0), street boy (0).

(*e*) Nurse (2), nurse maid (2), help (1).

(*f*) Carpenter (2), cabinet maker (1), carver (1), toy-maker (0).

The expression " one word " cannot be taken quite literally, otherwise " Post Mistress " (the form prescribed by the original examining authority) could not be accepted.

If a suitable adjective is used instead of a noun it should be considered correct. " Describe " suggests an adjective.

For two separate words given (two " shots " at the answer), one right and the other nearly right, (2). For two separate words, one right and the other wrong, (1).

APPENDIX II TO CHAPTER VI

TABLE 74

SPECIAL PLACE EXAMINATION (I) : ARITHMETIC, PART A
Total Marks for each fifth Candidate (Maximum = 40)

Candidate	Examiner A	B	C	D	E	F	G	H	J	K	Range
1	34	34	32	34	34	34	32	34	34	32	2
6	36	36	36	36	36	36	36	36	36	36	0
11	30	30	30	30	30	30	30	30	30	30	0
16	22	22	24	22	22	22	22	22	24	22	2
21	30	30	30	30	30	30	30	30	30	30	0
26	36	36	36	36	36	36	36	36	36	36	0
31	28	28	28	28	28	28	28	28	28	28	0
36	34	34	36	36	36	36	36	36	36	36	2
41	36	36	36	36	36	36	36	36	36	38	2
46	36	36	36	36	36	36	36	36	36	36	0
51	36	36	36	36	36	36	36	36	36	36	0
56	36	36	36	36	36	36	36	36	36	36	0
61	28	28	30	28	28	28	28	28	28	28	2
66	22	24	26	22	24	26	22	22	24	24	4
71	34	34	34	34	34	34	34	34	34	34	0
76	34	34	34	34	34	34	34	34	34	34	0
81	38	38	38	38	38	38	38	38	38	38	0
86	32	32	32	32	32	32	32	32	32	32	0
91	34	34	34	34	34	34	34	34	34	34	0
96	26	26	26	26	26	26	26	26	26	26	0
101	30	30	30	30	30	30	30	30	30	30	0
106	28	28	28	28	28	26	28	28	32	28	6
111	32	32	32	32	32	32	32	32	32	30	2
116	40	40	40	40	40	40	40	40	40	40	0
121	34	34	34	34	34	34	34	34	34	34	0
126	36	36	36	38	36	36	38	36	36	38	2
131	40	40	40	40	40	40	40	40	40	40	0
136	30	30	30	30	30	30	30	30	30	30	0
141	38	38	38	38	38	38	38	38	38	38	0
146	36	36	36	36	36	36	36	36	36	36	0

TABLE 75

SPECIAL PLACE EXAMINATION (I) : ARITHMETIC, PART B
Total Marks for each fifth Candidate (Maximum = 60)

Candidate	Examiner A	B	C	D	E	F	G	H	J	K	Range
1	39	22	22	22	16	30	22	26	26	22	23
6	29	34	37	30	29	26	30	34	34	26	11
11	26	22	26	26	26	26	26	26	26	26	4
16	36	41	24	36	44	33	21	33	31	36	23

TABLE 75—*continued*

Candidate	Examiner A	B	C	D	E	F	G	H	J	K	Range
21	31	31	39	31	31	31	26	31	31	35	13
26	15	15	15	15	15	15	15	26	15	15	11
31	22	27	19	22	30	30	26	19	30	26	11
36	49	46	49	42	49	49	11	49	46	44	38
41	27	27	27	27	27	27	22	27	27	27	5
46	33	29	26	30	37	37	41	38	38	29	15
51	22	22	22	22	22	22	33	22	22	25	11
56	32	19	29	24	32	32	15	21	24	17	17
61	27	15	15	15	24	15	38	15	15	15	23
66	46	41	45	46	57	45	38	41	39	41	19
71	22	22	22	22	22	22	17	22	22	22	5
76	49	49	42	49	44	49	39	49	34	34	15
81	57	57	57	46	57	51	37	57	51	57	20
86	28	25	34	24	24	21	15	21	33	24	19
91	22	35	19	31	34	31	31	31	31	19	16
96	15	15	15	15	15	15	15	15	15	15	0
101	34	26	34	30	34	30	26	30	30	30	8
106	26	30	39	30	26	34	26	26	43	26	17
111	29	34	31	47	39	51	36	34	34	34	22
116	25	25	25	25	37	45	18	25	49	10	39
121	42	42	42	46	42	42	18	42	42	42	28
126	57	57	57	54	57	57	52	57	57	57	5
131	44	44	44	44	52	60	49	52	52	44	16
136	22	22	22	22	22	22	26	19	22	22	7
141	41	39	34	36	41	52	30	41	49	24	28
146	45	37	34	37	37	34	30	30	22	37	23

TABLE 76

SPECIAL PLACE EXAMINATION (I): ENGLISH, PART A
Total Marks for each fifth Candidate (Maximum = 50)

Candidate	Examiner A	B	C	D	E	F	G	H	J	K	Range
1	31	27	31	41	29	29	23	21	14	23	27
6	27	27	29	25	17	27	31	25	17	31	14
11	33	31	39	33	37	39	23	27	21	29	18
16	31	29	45	33	39	41	37	25	21	25	24
21	35	31	31	45	33	41	33	33	31	23	22
26	27	19	37	37	39	29	23	29	17	31	22
31	35	29	43	43	43	39	35	35	25	21	22
36	37	29	29	35	35	27	23	27	15	21	22
41	25	21	27	21	23	21	25	23	19	21	8
46	43	31	49	33	33	39	39	37	27	31	22
51	29	38	27	35	31	33	27	27	24	31	14
56	25	25	45	27	47	33	35	33	11	31	36
61	29	37	33	31	37	29	23	21	19	25	18
66	23	15	29	13	25	21	13	17	3	15	26
71	29	27	19	27	29	27	27	25	25	21	10
76	37	40	45	47	47	41	43	41	35	35	12
81	45	46	47	47	47	41	41	47	39	43	8

TABLE 76—*continued*

Candidate	Examiner A	B	C	D	E	F	G	H	J	K	Range
86	33	37	31	35	45	29	27	37	27	31	18
91	21	27	33	23	43	33	21	35	17	21	26
96	39	35	41	37	49	39	45	43	33	31	18
101	19	31	29	43	33	34	23	18	33	19	25
106	33	27	21	31	41	37	29	21	17	17	24
111	19	27	23	17	45	29	27	21	17	13	32
116	27	29	37	37	27	29	27	31	25	23	14
121	22	13	27	31	37	31	17	22	4	23	33
126	27	33	29	33	45	37	25	23	23	25	22
131	29	31	33	21	43	37	35	37	19	27	24
136	23	27	25	41	25	21	25	20	8	33	33
141	19	25	27	21	39	35	29	33	17	25	22
146	15	25	17	25	23	20	13	9	12	15	16

TABLE 77

SPECIAL PLACE EXAMINATION (I) : ENGLISH, PART B
Total Marks for each fifth Candidate (Maximum = 50)

Candidate	Examiner A	B	C	D	E	F	G	H	J	K	Range
1	35	41	39	39	40	37	30	38	35	28	13
6	27	31	28	32	32	25	24	27	28	22	10
11	34	45	39	39	36	39	37	36	35	36	11
16	28	34	33	42	37	28	28	34	31	23	19
21	22	29	29	36	34	20	22	26	25	20	16
26	30	34	32	31	33	30	30	29	28	24	10
31	25	27	33	34	30	23	25	29	29	24	11
36	36	36	37	42	39	33	34	38	34	32	10
41	29	36	31	37	36	33	22	31	40	30	18
46	38	44	42	41	44	41	37	41	36	36	8
51	33	35	36	40	37	32	33	36	33	30	10
56	29	28	36	38	37	27	31	34	34	33	11
61	29	35	33	34	40	29	24	28	34	28	16
66	31	32	35	41	38	28	31	30	34	31	13
71	18	27	23	32	25	17	16	22	19	20	16
76	29	37	35	35	34	31	30	30	29	34	8
81	39	39	42	41	43	38	39	40	39	44	6
86	27	29	28	34	37	25	27	32	27	28	12
91	29	39	31	31	37	29	28	31	30	28	11
96	30	31	30	32	39	25	28	27	32	29	14
101	29	33	37	36	38	30	28	29	35	30	10
106	23	32	31	39	38	23	22	29	30	26	17
111	24	24	22	22	29	24	13	25	28	26	16
116	30	36	29	45	40	24	26	33	34	32	21
121	23	30	27	35	37	24	23	28	27	25	14
126	31	39	38	36	43	36	29	38	34	34	14
131	32	39	34	42	38	34	29	36	33	34	13
136	31	40	40	43	42	35	33	40	36	31	12
141	16	20	23	31	25	12	14	20	18	16	19
146	18	29	25	30	27	17	14	19	22	18	16

CHAPTER VII

SPECIAL PLACE EXAMINATION (II): MARKING OF ENGLISH ESSAY

246. *Character of the Examination Paper.*—The question-paper was as follows :—

> Choose *one* of the following subjects and write in accordance with the directions given :—
>
> (a) Write a letter to a friend who has asked : "What on earth do you do with yourself on a wet day in the holidays ?"
>
> (b) Where would you prefer to live, in the town or in the country ? Say why.
>
> (c) Many a London child has never been to the seaside. Describe the seaside for him.
>
> (d) Finish the following story : It was now clear that I had lost my way. Moreover it had grown so dark that I could not read my map. . . .

The time allowed was 30 minutes.

247. *Special Object of the Investigation.*—The special object of the investigation was to compare the discrepancies between the marks awarded by ten different examiners, all experienced in the marking of Special Place examination English Essay scripts, when such essays are marked on impression only, with the discrepancies which occur when they are marked in accordance with a detailed marking-scheme.

248. *Procedure.*—Ten examiners experienced in marking for Special Place examinations were appointed from the panels of several different authorities (other than the authority by which the scripts were supplied). It should be added that the scripts were supplied by a different authority from that which supplied the scripts for the previous investigation.

A draft detailed marking-scheme was drawn up by Dr. Ballard and was submitted for criticism to each of the examiners, together with typed copies of fifteen trial scripts. A considerable

number of criticisms of the original marking-scheme were received. The scheme was modified to meet the criticisms of the examiners, and answers on all doubtful points were furnished to them. A copy of the scheme in its final form is printed in the Appendix to this chapter (pp. 138–141).

249. Typed copies were then made of one hundred and fifty other scripts. Each examiner received not only a typed copy of each of the essays (on which it was possible for him to insert marks), but also the script itself, so that he could mark for handwriting. The marks of the original examiners were completely removed from the scripts.

250. The one hundred and fifty scripts were not specially selected. From the marks it is obvious that they vary in merit from " very poor " to " very good."

251. The following instructions were issued to the examiners :—

> (i) Scripts 1–75 are to be marked *by impression only*. It is of the essence of the investigation that, in marking these scripts, no attempt should be made by the examiner to conform to the scheme of marking set out under II below,[1] or to any scheme of the kind. Examiners *are particularly requested to mark scripts* 1–75 *before they mark scripts* 76–150.
>
> (ii) Scripts 76–150 are to be marked according to the amended marking-scheme set out under II below.[1]
>
> (iii) The maximum mark for all scripts is 100.

The examiners were supplied with the amended marking-scheme (see Appendix, p. 138) from which the following paragraph is extracted[2] :—

> Marks are to be allotted as follows :—
>
> | | Marks |
> |---|---|
> | (i) Quantity, Quality, and Control of Ideas | 50 |
> | (ii) Vocabulary | 15 |
> | (iii) Grammar and Punctuation | 15 |
> | (iv) Structure of Sentences | 10 |
> | (v) Spelling | 5 |
> | (vi) Handwriting | 5 |
> | Total | 100 |

252. The main object of this investigation, as stated in para. 247 above, was to compare the results obtained by marking a set of scripts by impression with the results obtained by marking another group of scripts of approximately the same calibre by means of a detailed scheme. Steps were taken, therefore, in

[1] See Appendix to Chapter VII, p. 138.

[2] Compare paras. 172 and 173.

the first instance to ensure that the selection from the original group of Nos. 1–75, designated later as Set 1, and of Nos. 76–150, designated later as Set 2, was approximately a random one. To test the equivalence of these two sets we reshuffled and then renumbered them 1–150 ; they were then marked by three examiners, designated as X, Y and Z, who were asked to mark the whole of the scripts on *any* one method at their choice, without being informed of the precise object of this particular investigation. X, Y and Z were all members of a panel of examiners for a Special Place examination, though not for this particular one.

253. Subsequently the marks allotted by X, Y and Z were regrouped according to the original numbers 1–75 and 76–150, and a summary of the results obtained is shown below :—

TABLE 78

AVERAGE MARKS

Examiner	Set 1	Set 2
X	58·2	58·3
Y	41·0	41·0
Z	55·3	55·3

DISTRIBUTION OF MARKS

Marks	Examiner X		Examiner Y		Examiner Z	
	Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
0– 7				1	1	
8–15			3		1	
16–23	2	1		2	1	1
24–31	3	1	6	9	3	5
32–39	2	5	17	11	5	8
40–47	13	12	34	32	9	8
48–55	13	15	13	17	16	17
56–63	12	16		2	17	15
64–71	14	8	2	1	11	10
72–79	7	5			6	4
80–87	5	6			2	3
88–95	4	5			2	3
96–100		1			1	1
Total	75	75	75	75	75	75

254. The Table shows that the two sets of scripts are approximately equivalent, since (a) each of the three examiners assigns the same average mark to Set 1 that he assigns to Set 2 ; and (b) each examiner distributes his marks for Set 1 in approximately the same way as he distributes them for Set 2.

255. We may feel confident, therefore, that any difference between the markings of the two sets by each of the individual examiners employed on the main investigation was due to the difference of method employed and not to a difference of intrinsic quality between the two sets.

256. The first and most striking results of the main investigation are given below :—

TABLE 79

AVERAGE MARKS

	Examiner [1]										Difference between highest and lowest averages
	A	B	C	E	G	K	L	M	N	P	
Set 1 (Marking by Impression)	49·0	43·7	59·4	31·8	44·6	47·5	51·2	40·0	46·2	41·7	27·6
Set 2 (Detailed Marking)	60·6	54·6	62·3	58·8	58·5	49·3	53·5	50·5	55·9	54·5	13·0
Difference	11·6	10·9	2·9	27·0	13·9	1·8	2·3	10·5	9·6	12·8	

In every case the average of the marks awarded to Set 2, the scripts marked by details, is greater than the average of the marks awarded to Set 1, the scripts marked by impression. With Examiner E the difference of the averages is 27 marks ; with Examiners A, B, G, M, N and P, the difference is about 10 marks ; and with only three examiners, C, K and L, is the difference small—about 2 or 3 marks. We can confidently say that marking by details produces higher marks on the average than marking by impression.

257. The highest average mark by impression is 59·4 (Examiner C) and the lowest is 31·8 (Examiner E), a difference of 27·6 marks. The highest and lowest averages of the marks according to the detailed scheme are 62·3 (Examiner C) and 49·3 (Examiner K), a difference of 13·0 marks. The mean deviation of the averages of the impression marks is 5·2, and that of the averages of the detailed marks is 3·4. Thus the averages of the several examiners are closer to one another when the marking is made in accordance with the detailed scheme than when it is made by impression only.

[1] Examiners A, B, C, E, G and K are the examiners in English who were designated by those letters in the previous investigation on a Special Place examination. L, M, N and P are examiners who did not take part in the previous investigation, but, like the other examiners, they are all experienced in examining of this kind.

258.[1] The average range of marks was 36·5 for the marking by impression (Set 1) and 28·9 for the marking by a detailed scheme (Set 2); in other words, this method of analysis shows again that the marking by a detailed scheme yields on the whole closer results for the different examiners than the marking by impression.

In the marking by impression the highest range was 63, shown in the marks for Candidate No. 41, who received the following: 50, 63, 69, 15, 78, 62, 75, 48, 71, 64; and the lowest range was 13, in the case of Candidate No. 38, who received the following marks: 12, 12, 20, 12, 15, 10, 7, 10, 12, 18. In the marking by details, the highest range, 52, was shown in the case of Candidate No. 85, who received the following marks: 43·5, 26, 50, 42, 50, 58, 51, 57, 78, 39; and the lowest range was 14·5, for Candidate No. 141, who received the following marks: 72·5, 69, 67, 62, 59, 58, 67, 61, 70, 63.

[1] The following Tables give the marks and ranges for every fifth candidate, and are inserted so as to convey to the reader the kind of differences between the different examiners. :—

Special Place English Essay: *marking by impression only*

	Examiner										
Cand.	A	B	C	E	G	K	L	M	N	P	Range
1	25	15	45	40	40	28	46	35	35	35	31
6	65	64	59	33	52	65	61	52	60	50	32
11	65	88	90	60	92	85	77	60	63	70	32
16	30	10	55	17	28	36	28	15	35	25	45
21	40	49	68	25	34	45	58	52	50	41	43
26	62	43	58	30	44	58	55	31	43	34	32
31	48	44	64	55	40	48	60	58	45	35	29
36	43	25	52	17	33	52	34	20	18	36	35
41	50	63	69	15	78	62	75	48	71	64	63
46	50	68	78	42	58	45	72	47	65	50	36
51	55	52	54	35	45	42	44	36	43	33	22
56	18	15	54	15	17	20	11	10	12	17	44
61	45	47	60	17	40	38	58	56	52	42	43
66	40	45	58	15	25	40	35	32	43	30	43
71	39	35	52	32	40	38	40	25	41	28	27

Special Place English Essay: *marking by details*

	Examiner										
Cand.	A	B	C	E	G	K	L	M	N	P	Range
76	65	63	64	67	68	65	53	52	49	57	19
81	42½	46	43	40	62	20	38	36	60	39	42
86	45	42	54½	64	62	45	35	44	42	38	29
91	68½	64	69½	76	79	65	65	59	64	70	20
96	56	56	70	58	80	48	44	31	43	49	49
101	70½	89	89	73	94	74	86	83	76	87	21
106	78	79	98	88	87	92	88	88	87	89	20
111	62	59	55½	44	49	41	61	55	70	59	29
116	71	64	64	58	62	66	58	50	55	52	21
121	79	91	82	70	80	51	60	67	64	56	40
126	99	88	68	68	89	77	70	71	90	90	31
131	44	26	55½	32	40	44	35	25	48	35	30½
136	51½	40	43	42	31	24	28	29	59	39	35
141	72½	69	67	62	59	58	67	61	70	63	14½
146	46½	37	49½	44	39	34	28	24	32	38	25½

259. The Table below gives the distribution of the ranges :—

TABLE 80

DISTRIBUTION OF RANGES

Range	Set 1	Set 2
10–14	1	1
15–19		7
20–24	3	13
25–29	10	19
30–34	19	19
35–39	16	7
40–44	16	8
45–49	4	
50–54	3	1
55–59	2	
60–64	1	
Total	75	75

The Table brings out in another way the fact that the marks awarded by the examiners to an individual script were (as was shown by the average ranges) closer together when the examiners used the detailed scheme than when they marked by impression, though in neither case was any script awarded the same mark by all the examiners.

260. The range of marks for a given script may be high, partly because of differences of standard of the examiners' marks, and partly on account of the presence of " random elements " in their marking. We have observed that the range of marks is on the whole higher with marking by impression than with detailed marking. But the method of analysis described in Part II (para. 536 below) indicates that the greater differences shown in the marking by impression (see para. 257 above) are not due to a higher figure for the random marking, but to a greater difference between the standards adopted by the different examiners. The analysis shows that the element of random marking has roughly speaking the same magnitude in both cases.

261. This last point is important. It means that the use of a detailed marking-scheme does conduce to a closer approximation of the standards of the different examiners, but, on the other hand, that it does nothing to reduce the element of random marking. It is of course well known that manipulations in the office of the examination authority may be used to make the marks of different examiners tend to conform to a common standard ; the use of a detailed marking-scheme attains the same object in a different way, but nothing more, if the final marks alone are taken into consideration.

262. The main object of the investigation has now been achieved, since we have ascertained that there are greater discrepancies between marks awarded by impression than between marks awarded by details. Moreover, as will be shown in Part II, on the basis of the theory put forward therein, it appears that these discrepancies are entirely due to greater differences in the standards of marking of different examiners when they mark by impression. There is no appreciable difference between the random variations of the examiners in the two methods of marking.

263. *Detailed Investigation of Discrepancies.*—The following Table shows clearly the differences between the distribution of marks awarded by the examiners to the two sets of scripts, denoted in the Table by the figures (1) and (2)[1] :—

TABLE 81

DISTRIBUTION OF MARKS

Marks	Examiner A		B		C		E		G		K		L		M		N		P	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
0-7			1										1							
8-15	3		6	2			9		3		1	1	2	1	5	1	3		1	
16-23	2		6	1	1		23		7	1	3	2			5	2	4	1	4	
24-31	8	1	6	6	1		10	3	6	2	4	4	5	6	11	9	6	4	13	
32-39	7	3	5	9			8	2	16	4	10	13	10	11	17	9	12	5	27	11
40-47	14	12	19	8	4	10	11	19	18	8	21	17	14	11	13	14	20	15	7	12
48-55	15	13	13	12	20	16	6	9	8	17	18	12	13	13	10	11	7	13	10	15
56-63	9	14	8	12	25	18	5	12	3	17	8	12	13	12	8	11	10	13	3	16
64-71	11	15	7	11	18	16	3	13	5	13	4	7	5	10	5	12	9	12	6	7
72-79	3	11	3	5	4	7		9	3	5	4	4	8	5		1	4	5	2	5
80-87	1	4		4		3		4	4	6	2	2	4	3	1	3		4	2	3
88-95	2		1	5	2	3		4	1	2		1		3		2		3		2
96-100		2				2			1											1

We have already seen (para. 256 above) that the averages of the different examiners differ. It is interesting to compare for each examiner the number of scripts to which he has assigned less than 40 marks, and the number to which he has assigned 72 marks or more.

	Examiner A		B		C		E		G		K		L		M		N		P	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
No. of scripts to which less than 40 marks are allotted	20	4	24	18	2	0	50	5	32	7	18	20	18	18	38	21	25	10	45	14
No. of scripts to which 72 marks or more are allotted	6	17	4	14	6	15	0	17	9	13	6	7	12	11	1	6	4	12	4	11

[1] (1) Set 1 ; Scripts Nos. 1–75—marked by impression.
(2) Set 2 ; Scripts Nos. 76–150—marked by details.

264. The Table brings out two features in the marking, (a) the difference between the results of the different examiners marking by the same method; (b) the difference between the results of the same examiner marking by different methods.

Thus, in the marking by impression, Examiner E awards fifty marks of less than 40, and Examiner C only two; on the other hand, Examiner L gives twelve marks of 72 or more, and Examiner E gives none.

Again, in the marking by details, Examiner M gives twenty-one marks of less than 40 and Examiner C gives none; while Examiners A and E give seventeen marks of 72 or more and Examiner M gives only six.

K and L are the only two examiners whose marks show approximately similar distributions, and whose averages are approximately the same when marking by the two different methods (see para. 256 above).

265. The following are the standard deviations of the distributions in Table 81, indicating the amount of variability amongst the candidates :—

	A	B	C	E	G	K	L	M	N	P	Average
(1)	17·4	18·2	10·8	15·8	19·3	14·3	17·2	15·5	16·0	15·4	16·0
(2)	14·4	19·3	13·3	15·8	14·7	15·6	17·1	17·2	16·1	16·2	16·0

The averages of the two standard deviations for the two methods of marking are thus the same; in other words, the method of marking by impression and the method of marking by details produce on the average the same degree of discrimination between the merits of the different candidates, the same "spread" of the marks.

With three examiners, A, C and G, the difference between the standard deviations of their markings by the two methods is notable; A and G are able to distribute their marks over a greater range when marking by impression; C, on the contrary, distributes them more when marking by details. With the other examiners the difference of method in marking does not seem to affect their powers of discrimination to an appreciable extent.

---

266. We give below in Table 82 the distributions of the marks of the several examiners for the various elements marked separately.

TABLE 82

Distribution of Marks Awarded for the Various Elements

Quantity, Quality, and Control of Ideas

(Maximum = 50)

Marks	Examiner									
	A	B	C	E	G	K	L	M	N	P
1–5		2				7	1	5	3	
6–10	4	10			4	6	3	9	12	4
11–15	5	8		13	3	18	9	16	8	6
16–20	11	10	16	16	11	7	10	12	15	15
21–25	14	17	5	12	22	15	10	9	13	18
26–30	16	6	31	16	16	11	19	14	10	8
31–35	13	9	5	8	8	8	12	4	7	17
36–40	6	5	11	7	7	2	6	1	6	
41–45	4	8	4	3	4	1	5	4	1	6
46–50	2		3					1		1

Vocabulary

(Maximum = 15)

Marks	Examiner									
	A	B	C	E	G	K	L	M	N	P
1					1			1		
2					3	3			2	
3		2			7	8	2	4	3	2
4	2	1		1	12	9	5	8	3	2
5	6	5		3	7	10	8	6	7	7
6	8	4	1	5	10	6	6	7	9	10
7	7	7	7	12	9	11	17	3	9	18
8	22	14	29	15	12	13	9	8	10	12
9	9	7	14	8	3	7	12	9	13	8
10	7	15	10	11	3	3	4	11	5	7
11	5	4	3	5	3		4	5	4	2
12	3	8	2	6	1	2	2	9	3	
13	3	2	1	1	2	1	3	3	5	3
14		4	3	4	1	1	3	1		3
15	3	2	5	4	1	1			2	1

Grammar and Punctuation

(Maximum = 15)

Marks	Examiner									
	A	B	C	E	G	K	L	M	N	P
0		2				2				1
1						1	1	1		
2	1	4					1	1		2
3	3	3	1	2		3	6	3		2
4	2	2		6			4	5		4
5	1	2		7		5	11	10		5
6	1	5	1	4			9	8		7
7	2	9	4	7		13	11	5	5	2
8	5	12	2	13	2		9	10	1	5
9	7	7	13	5	4	17	8	7	14	17
10	8	12	6	8	3		7	11	3	8
11	5	3	21	5	6	10	2	8	25	6
12	12	6		7				4	8	7
13	12	5	16	6	19	14	4	2	10	7
14	11	3		2	4		2		4	1
15	5		11	3	37	10			5	1

Structure of Sentences

(Maximum = 10)

Marks	Examiner									
	A	B	C	E	G	K	L	M	N	P
0		2	6		9	1	1			2
1	1	1			2	3		1	1	1
2	1	4	28	4	13	15	2	2	2	5
3	5	4	4	3	10	14	14	2	1	7
4	14	6	12	6	18	12	17	18	14	19
5	20	7	5	14	7	12	12	15	7	14
6	14	13	7	6	6	8	14	17	10	12
7	8	16	2	11	4	6	6	9	15	7
8	6	11	7	19	3	1	3	9	14	2
9	3	11		9	2		5	2	8	5
10	3		4	3	1	3	1		3	1

Spelling

(Maximum = 5)

Marks	Examiner									
	A	B	C	E	G	K	L	M	N	P
0		2				1	1			
1	1	3	1	3	1	1	5	3	2	
2	1	13	5	4	2	5	22	8	4	7
2½	4		1		2					
3	2	15	10	9	7	18	26	21	9	21
3½	13		5		1					
4	23	28	18	17	26	28	21	27	30	27
4½	15		11		12			1		
5	16	14	24	42	24	22		15	30	20

Handwriting

(Maximum = 5)

Marks	Examiner									
	A	B	C	E	G	K	L	M	N	P
1		1								4
2		3	1							4
3	2	20	6					1		9
4	5	42	33	4	1	3	2	7		15
4½								1		
5	68	9	35	71	74	72	73	66	75	43

267. An analysis of the marks allotted by the different examiners for the different elements in the Scheme is interesting. We shall, in referring to them, use the following abbreviated nomenclature :—

For Quantity, Quality, and Control of Ideas we shall use " Ideas "
,, Vocabulary ,, ,, ,, Vocabulary
,, Grammar and Punctuation ,, ,, ,, " Grammar "
,, Structure of Sentences ,, ,, ,, " Structure "
,, Spelling ,, ,, ,, Spelling
,, Handwriting ,, ,, ,, Handwriting.

268. Table 82 above gives in detail the distribution of marks awarded for these different elements, and Tables 83 and 84 below give respectively the averages obtained from Table 82 and these averages expressed as percentages of the relevant maximum in each case :—

TABLE 83

AVERAGE MARKS AWARDED BY THE EXAMINERS

	Maximum Marks	Examiner A	B	C	E	G	K	L	M	N	P
" Ideas "	50	26·8	24·0	29·6	25·2	25·7	20·2	25·6	20·6	21·3	25·0
Vocabulary	15	8·5	9·1	9·3	9·2	6·5	6·6	7·9	8·1	8·0	7·9
" Grammar "	15	10·8	8·3	11·1	8·8	13·4	9·6	7·2	7·7	11·1	8·7
" Structure "	10	5·6	6·1	3·9	6·5	3·7	4·1	5·1	5·5	6·4	4·9
Spelling	5	4·1	3·4	4·0	4·2	4·2	3·8	2·8	3·6	4·1	3·8
Handwriting	5	4·9	3·7	4·4	4·9	5·0	5·0	5·0	4·9	5·0	4·2
Total		60·6	54·6	62·3	58·8	58·5	49·3	53·5	50·5	55·9	54·5

TABLE 84

AVERAGE MARKS AWARDED BY THE EXAMINERS EXPRESSED AS PERCENTAGES OF THE RELEVANT MAXIMUM IN EACH CASE

	Maximum Marks	Examiner A	B	C	E	G	K	L	M	N	P
" Ideas "	50	53·6	48·0	59·2	50·4	51·4	40·4	51·2	41·2	42·6	50·0
Vocabulary	15	56·6	60·7	62·0	61·3	43·3	44·0	52·6	54·0	53·3	52·6
" Grammar "	15	72·0	55·3	74·0	58·7	89·3	64·0	48·0	51·6	74·0	58·0
" Structure "	10	56·0	61·0	39·0	65·0	37·0	41·0	50·5	55·0	64·0	49·0
Spelling	5	81·0	68·0	80·0	84·0	84·0	76·0	56·0	72·0	82·0	76·0
Handwriting	5	98·0	74·0	88·0	98·0	100	100	100	98·0	100	84·0

The average marks awarded for " Ideas " (maximum 50) range from 20·2 (K) to 29·6 (C), with a difference equivalent to 19% of the maximum ; those for Vocabulary (maximum 15) range from 6·5 (G) to 9·3 (C), with a difference equivalent to 19% ; those for " Grammar " (maximum 15) from 7·2 (L) to

13·4 (G), with a difference equivalent to 41% ; those for " Structure " (maximum 10) from 3·7 (G) to 6·5 (E), with a difference equivalent to 28% ; those for Spelling (maximum 5) from 2·8 (L) to 4·2 (E, G), with a difference equivalent to 28% ; and those for Handwriting (maximum 5) from 3·7 (B) to 5·0 (G, K, L, N), with a difference equivalent to 26%.

269. The differences in standards of marking for these various elements are apparent. They exist to a considerable degree in marking " Grammar," and even in marking Spelling and Handwriting. Examiner B differs fundamentally from the others in his view of what may be expected from children in the way of passable handwriting ; and Examiner L clearly has very different ideas from most of the other examiners on the subject of spelling.

270. We might have expected that most agreement would have been reached in respect of the two minor elements, Spelling and Handwriting, for each of which 5 marks were obtainable, but, as the averages show, even in the estimation of these elements there are wide differences.

271. Curiously enough, the marks of the different examiners for " Ideas " and Vocabulary resemble each other more closely, as far as can be judged from averages, than those for Spelling and Handwriting.

272. In order to see whether the differences of standard between the different examiners vary with the element in composition which they are estimating, or remain constant in judging the different elements, we have, in Table 85 below, placed the examiners in order according to the average marks which they have given to the whole seventy-five candidates for each element. The examiners with the highest averages have been placed first in order.

TABLE 85

EXAMINERS IN THE ORDER OF THE AVERAGE MARKS ASSIGNED BY THEM TO THE DIFFERENT ELEMENTS

Examiner	" Ideas "	Vocabulary	" Grammar "	" Structure "	Spelling
A	2	4	4	4	3½
B	7	3	8	3	9
C	1	1	2½	9	5
E	5	2	6	1	1½
G	3	10	1	10	1½
K	10	9	5	8	6½
L	4	7½	10	6	10
M	9	5	9	5	8
N	8	6	2½	2	3½
P	6	7½	7	7	6½

This Table corresponds to Table 52 on p. 89.

It is obvious that whereas some examiners, e.g. A and P, keep approximately the same places in the list for all the elements, others, especially G and C, sometimes mark high and sometimes mark low.

273. Table 86 below gives the mean deviations of the marks awarded for the different elements considered, expressed as percentages of the relevant maximum in each case. The mean deviations are measures of the variability of the candidates in respect of these elements.

TABLE 86

MEAN DEVIATIONS OF THE MARKS AWARDED FOR THE DIFFERENT ELEMENTS EXPRESSED AS PERCENTAGES OF THE MAXIMUM FOR EACH ELEMENT

	Examiner									
	A	B	C	E	G	K	L	M	N	P
"Ideas"	15·2	18·6	12·8	14·6	13·4	16·8	15·6	17·6	16·8	15·2
Vocabulary	12·1	14·7	11·3	14·0	16·0	14·7	14·0	17·3	15·3	12·7
"Grammar"	17·3	17·3	12·7	18·0	10·7	20·7	15·3	16·0	10·0	16·7
"Structure"	15	18	21	18	18	17	16	14	17	16
Spelling	12	20	16	18	12	18	16	18	14	16
Handwriting	4	12	12	2	0	2	1	4	0	18

The Table also affords an indication of the relative powers of different examiners to discriminate between the performances of the different candidates in regard to the same elements. (Thus, in regard to "Ideas," Examiner B, with a mean deviation of 18·6, "spreads" his marks more than Examiner C, whose mean deviation is 12·8.) It also indicates the relative powers of the same examiners to discriminate in dealing with different elements. The most striking feature in the Table is the similarity in the spreading of the marks in all subjects except that of Handwriting. The special direction given in regard to Handwriting (see p. 141) sufficiently explains the smallness of the deviations in the marks allotted by the majority of the examiners for this element.

The contrast between the results for Handwriting and for Spelling is remarkable. It would appear that legibility of handwriting among these children is fairly constant, whereas spelling is not.

274. The following Table shows the order of the examiners in respect of their mean deviations, those with the largest mean deviation being placed first.

TABLE 87

EXAMINERS IN THE ORDER OF THE MEAN DEVIATIONS OF THEIR MARKS FOR THE DIFFERENT ELEMENTS

Examiner	"Ideas"	Vocabulary	"Grammar"	"Structure"
A	6½	9	3½	9
B	1	4½	3½	3
C	10	10	8	1
E	8	6½	2	3
G	9	2	9	3
K	3½	4½	1	5½
L	5	6½	7	7½
M	2	1	6	10
N	3½	3	10	5½
P	6½	8	5	7½

Compare para. 188 *et seq.* above.

275. We return to Table 82 in order to point out certain idiosyncrasies of the examiners relating to the different elements in the examination.

"*Ideas.*"—The number of marks of 36 or more varies from 3 (Examiner K) to 18 (Examiner C). Examiner C never gives less than 16 marks, while five examiners, B, K, L, M, N, award marks below 6.

Vocabulary.—Here Examiner C never gives less than 6 marks ; but, on the other hand, G and M go as low as one mark, and other examiners frequently give 2, 3, 4 or 5.

"*Grammar.*"—Examiner G gives the full marks to thirty-seven candidates, while B, L and M never award the maximum. Examiner G never gives less than half marks ; N never gives less than 7, while other examiners award zero to certain scripts. The views of the examiners seem irreconcilable on these points. Examiner K only uses the odd marks 1, 3, 5, 7, etc. The Table shows that C has a tendency to the same habit ; and this is true also of his marks for "Ideas" and Spelling.

Spelling.—Examiner E gives full marks to forty-two candidates, while Examiner L does not give full marks to anyone. B, K and L all give zero marks to some candidates.

Handwriting.—Examiner N gives full marks to all the candidates, and E, G, K and L give full marks to the vast majority. But B and P spread their marks ; and B gives the mark 4 to more than half the candidates.

276. Whereas, as we have seen, Examiners C and K in some cases have twice as many marks at their disposal as they need, A, G and M do not find enough scope in the marks suggested and introduce half-marks. C also introduces half-marks.

277. Table 88 below gives particulars of the ranges (differences

between the highest and lowest marks assigned to the various scripts) for the different elements evaluated :—

TABLE 88

"Ideas" (Maximum 50)		Vocabulary (Maximum 15)		"Grammar" (Maximum 15)		"Structure" (Maximum 10)		Spelling (Maximum 5)		Handwriting (Maximum 5)	
Range	No. of Scripts	Range	No. of Scripts	Range	No. of Scripts	Range	No. of Scripts	Range	No. of Scripts	Range	No. of Scripts
7	1	2	1	2	1	2	3	1	16	0	8
10	1	3	8	4	4	3	9	1½	3	1	36
12	2	4	10	5	3	4	15	2	35	2	22
13	1	5	22	6	10	5	22	2½	1	3	5
14	6	6	19	7	9	6	18	3	16	4	4
15	11	7	5	8	14	7	6	3½	1		
16	5	8	6	9	14	8	2	4	3		
17	6	10	3	10	8						
18	5	11	1	11	9						
19	6			12	1						
20	1			13	2						
21	8										
22	4										
23	4										
24	2										
25	5										
26	1										
27	2										
28	2										
31	1										
33	1										

To take two examples, there were eleven scripts for which the greatest difference between the marks assigned in respect of "Ideas" was 15 marks (the maximum being 50); and in respect of Vocabulary there were twenty-two scripts for which the corresponding difference was 5 marks (the maximum being 15).

278. The following are the average ranges obtained from Table 88 :—

	"Ideas"	Vocabulary	"Grammar"	"Structure"	Spelling	Handwriting
Maximum	50	15	15	10	5	5
Average Range	19·1	5·5	8·1	4·9	2·1	1·5
Average Range as % of maximum	38	37	54	49	42	30

Thus the average difference between the extreme marks awarded is a high percentage of the maximum mark in each case.

279. It will be seen that the greatest average range, expressed as a percentage of the maximum, occurs in "Grammar," and the least in Handwriting. It will be interesting to examine the ranges for each element in somewhat greater detail (see Table 88).

"*Ideas*" (Maximum 50). The highest range is 33; Candidate No. 98 receives marks varying from 10 to 43; the lowest range is 7; Candidate No. 91 receives marks varying from 24 to 31. For twelve candidates the range is 25 or more.

Vocabulary (Maximum 15). The highest range is 11; Candidate No. 142 receives marks varying from 4 to 15. The lowest

range is 2 ; Candidate No. 139 receives marks varying from 7 to 9. For ten candidates the range is 8 or more.

" *Grammar* " (Maximum 15). The highest range is 13. Candidate No. 85 and Candidate No. 111 receive marks varying from 2 to 15 and from 0 to 13 respectively. The lowest range is 2 ; Candidate No. 129 receives marks varying from 13 to 15. For forty-eight candidates the range is 8 or more.

" *Structure* " (Maximum 10). The highest range is 8 ; Candidates Nos. 95 and 124 receive marks varying from 2 to 10 and from 0 to 8 respectively. The lowest range is 2 ; Candidates Nos. 129, 131 and 148 receive marks varying from 8 to 10, from 2 to 4, and from 2 to 4 respectively. For forty-eight candidates the range is 5 or more.

Spelling (Maximum 5). The highest range is 4 ; Candidates Nos. 93, 138 and 147 receive marks varying from 0 to 4, 1 to 5, and 1 to 5 respectively. The range is only 1 mark in sixteen cases ; for twenty-one candidates the range is $2\frac{1}{2}$ or more.

Handwriting (Maximum 5). This is the only element in which we find cases of complete agreement among the examiners ; there are eight cases in which such complete agreement occurs. The highest range is 4, for Candidates Nos. 100, 113, 148 and 149 ; the marks in each case vary from 1 to 5.

280. It will be seen that there are quite large numbers of candidates for whom the ranges of marks are as great as half the maximum, in respect of all the elements of the test except Handwriting. The large differences between the marks awarded by the several examiners to a script are of course a necessary corollary to the differences of standard as shown by the differences of their average marks (see Table 83 above).

---

281. The size of the range may be specially influenced by one or two examiners who mark exceptionally high or low as the case may be. We have therefore constructed Table 89 below to show to what extent there is agreement between the different examiners in respect of the different elements. In order to simplify the presentation the marks for each element have been reduced to five grades, in accordance with the following scheme :—

		GRADES				
		I	II	III	IV	V
" Ideas "	Marks	3–9	10–19	20–29	30–39	40–50
Vocabulary	,,	1–3	4–6	7–9	10–12	13–15
" Grammar "	,,	0–3	4–6	7–9	10–12	13–15
" Structure "	,,	0–1	2–3	4–5	6–7	8–10
Spelling	,,	0–1	2, $2\frac{1}{2}$	3, $3\frac{1}{2}$	4, $4\frac{1}{2}$	5
Handwriting	,,	1	2	3	4	5

Marks falling within the limits stated have been treated as the same. We have the following results[1] :—

TABLE 89

NUMBER OF CASES OF AGREEMENT AMONGST EXAMINERS

Examiners agreeing	"Ideas"	Vocabulary	"Grammar"	"Structure"	Spelling	Handwriting
10		1	1	1		8
9, 1	1	3		3	5	22
8, 2	4	4	1	2	8	9
8, 1, 1	1	2		2	7	6
7, 3	6	6	1	3	2	7
7, 2, 1	7	9	1	2	8	10
7, 1, 1, 1	1	2		2		1
6, 4	7	7	4	3	7	2
6, 3, 1	6	5	3	4	5	2
6, 2, 2	3	4	4	5	3	1
6, 2, 1, 1	3	1	4	1	3	1
6, 1, 1, 1, 1			1			
5, 5	7	2	1	1		
5, 4, 1	15	15	7	4	4	
5, 3, 2	3	5	4	11	7	1
5, 3, 1, 1		2	4	4	3	1
5, 2, 2, 1	2		4	4	3	1
5, 2, 1, 1, 1						1
4, 4, 2	1	2		1	3	1
4, 4, 1, 1	1	1	3	2	1	
4, 3, 3	2	3	4	7	1	
4, 3, 2, 1	3	1	14	6	4	1
4, 3, 1, 1, 1				1		
4, 2, 2, 2			3	3		
3, 3, 3, 1	1		4	1		
3, 3, 2, 2	1		3	2	1	
3, 3, 2, 1, 1			3			
3, 2, 2, 2, 1			1			
10 agree		1	1	1		8
9 ,,	1	3		3	5	22
8 ,,	5	6	1	4	15	15
7 ,,	14	17	2	7	10	18
6 ,,	19	17	16	13	18	6
5 ,,	27	24	20	24	17	4
Rest	9	7	35	23	10	2
6 or more agree	39	44	20	28	48	69
Rest	36	31	55	47	27	6

In Table 89 above the figures in the left-hand column show the extent of agreement as to these new grades among the

[1] Compare Table 47.

examiners. Thus, 7, 2, 1 means that seven examiners agree on one grade, two examiners agree on another and the remaining examiner gives a mark which indicates still another grade. Similarly, 4, 3, 1, 1, 1 means that all five grades are assigned to a script, three of them each by a single examiner, one by four examiners, and the fifth by three examiners.

282. Even when we reduce the number of grades in this way, we see that there is for the most part a conspicuous absence of agreement among the examiners.

Taking the categories Vocabulary, "Grammar" and "Structure" we find for each only a single case in which there is complete agreement as to grade among the examiners, and very few cases in which nine examiners out of ten agree.

The number of cases in which six or more examiners out of the ten agree is not large, it being remembered that the total number is 75. They are as follows : " Ideas," 39 ; Vocabulary, 44 ; " Grammar," 20 ; " Structure," 28 ; Spelling, 48 ; Handwriting, 69.

283. Thus the greatest agreement is reached in regard to Handwriting, Spelling and Vocabulary, and the least in regard to " Grammar " and " Structure." The results indicated by the investigation of the ranges or maximum differences are thus confirmed by the investigation of the cases of agreement.

284. It may be pointed out that figures of the same kind were obtained in the earlier investigation of a Special Place examination (see paras. 175-177 above).

---

285. We saw in para. 256 above that marking by details yielded higher marks on the average than marking by impression. It seemed of interest to ascertain if possible whether the higher marking affects all the scripts, good or bad, in equal measure, or whether the difference is itself dependent on the quality of a script.

286. Since the same scripts were not marked twice, first by impression and then by the detailed scheme, we must first find some criterion of the relative merits of the scripts before we can solve the question raised in the preceding paragraph, and this we can obtain from the independent markings of Examiners X, Y, and Z, who did not take part in the major investigation (see paras. 252 and 253 above). We have seen that each of these examiners assigned approximately the same average marks to the scripts 1–75 and 76–150 ; but this does not give us a sufficient comparison of the quality of individual scripts in the two sets.

The following method was therefore adopted for obtaining such a comparison.

287. We arranged (a) the scripts 1–75 (Set 1), and (b) the scripts 76–150 (Set 2) in the ascending order of the marks assigned to them by Examiner X and divided each set into five groups indicated by the adjectives "Poor," "Fair," "Moderate," "Good," and "Very Good," as set out in the following Table. The marks assigned to each script of Set 1 and those assigned to each script of Set 2 are shown in separate columns; the differences between the marks of the corresponding scripts are given.

TABLE 90

POOR			FAIR			MODERATE			GOOD			VERY GOOD		
Set 1 Marks	Set 2 Marks	Difference	Set 1 Marks	Set 2 Marks	Difference	Set 1 Marks	Set 2 Marks	Difference	Set 1 Marks	Set 2 Marks	Difference	Set 1 Marks	Set 2 Marks	Difference
18	16	−2	44	46	+2	52	52	0	64	60	−4	72	74	+2
18	30	+12	46	46	0	52	54	+2	64	62	−2	74	74	0
28	32	+4	46	46	0	54	54	0	64	62	−2	76	74	−2
28	34	+6	46	46	0	56	54	−2	64	62	−2	78	80	+2
30	34	+4	46	48	+2	56	56	0	66	62	−4	78	80	+2
36	36	0	48	48	0	56	56	0	66	64	−2	78	80	+2
38	36	−2	48	48	0	58	56	−2	66	64	−2	80	82	+2
40	40	0	48	50	+2	58	56	−2	66	64	−2	80	86	+6
40	40	0	50	50	0	60	58	−2	66	64	−2	84	86	+2
42	40	−2	50	50	0	60	58	−2	68	64	−4	86	88	+2
42	42	0	50	50	0	60	58	−2	68	66	−2	86	88	+2
42	44	+2	50	50	0	60	58	−2	68	68	0	90	90	0
44	44	0	50	50	0	62	58	−4	70	68	−2	90	92	+2
44	44	0	50	52	+2	62	60	−2	70	72	+2	90	92	+2
44	46	+2	52	52	0	62	60	−2	72	72	0	94	96	+2
Averages		+1·60			+0·53			−1·33			−1·87			+1·73

288. From the above Table we see that the lowest mark assigned by X to Set 1 was 18 (for two scripts) and the highest mark was 94; whereas for Set 2 his lowest mark was 16 and his highest 96. We also see that there is a good deal of agreement between the marks assigned the same rank in the two ordered Sets. The differences between the two scripts occupying the same rank are marked with a + sign when the mark of the script of Set 2 is higher than that of Set 1, and a − sign when it is lower.

289. Instead of dealing with individual scripts, we shall now set down the differences between the *average marks* in each group, i.e. between those belonging to Set 1 and those belonging to Set 2. The marks of Examiners Y and Z have been treated in the same way as those of X. Table 91 below shows the results of the three examiners analysed in this way.

TABLE 91

DIFFERENCES BETWEEN AVERAGES OF CORRESPONDING GROUPS OF SET 1 AND SET 2

(When the mark for Set 2 is greater than for Set 1 the mark is prefixed by a plus sign; when it is lower, by a minus sign)

		Examiner X	Examiner Y	Examiner Z
Group	Poor	+1·60	−1·20	+1·20
	Fair	+0·53	+0·40	−2·33
	Moderate	−1·33	−0·53	−0·13
	Good	−1·87	+1·53	−0·40
	Very Good	+1·73	−0·33	+1·87
	All Scripts	+0·13	−0·03	+0·04

The differences of the averages for the scripts are, as we have seen, almost negligible, but the averages of the Groups differ by greater amounts.

290. Since, on the whole, Examiner X found that Set 2 was superior to Set 1 by 0·13 marks, we now subtract this from each of the differences, and, treating the differences of Examiners Y and Z in the same way, we get the slightly corrected Table 92 :—

TABLE 92

AVERAGE DIFFERENCES OF MARKS ALLOTTED BY EXAMINERS X, Y, AND Z TO CORRESPONDING GROUPS OF SET 1 AND SET 2 AFTER REDUCTION OF THE MARKS FOR EACH SET AS A WHOLE TO THE SAME STANDARD

		Examiner X	Examiner Y	Examiner Z	Average
Group	Poor	+1·47	−1·17	+1·16	+0·49
	Fair	+0·40	+0·43	−2·37	−0·51
	Moderate	−1·46	−0·50	−0·17	−0·71
	Good	−2·00	+1·56	−0·44	−0·29
	Very Good	+1·60	−0·30	+1·83	+1·04

291. While there is not an exact agreement between X, Y, and Z as to the relative merits of the corresponding groups, the differences are small and on the average do not exceed 1 mark. We conclude that if we split up the two Sets each into five Groups in this way, we can legitimately consider the corresponding groups in the two Sets as approximately equivalent in merit, and hence that if the marks of Examiners A to P are *not* the same for these corresponding Groups, the difference is due to the difference in the method of marking, Set 1 having been marked by impression and Set 2 by details.

292. Let us now examine these differences, shown in Table 93 below :—

TABLE 93

AVERAGE DIFFERENCES BETWEEN DETAILED AND IMPRESSION MARKS GROUP BY GROUP

Group	Examiner A	B	C	E	G	K	L	M	N	P
Poor	+17·83	+10·73	+0·80	+23·80	+17·73	−0·13	+3·87	+8·53	+11·33	+9·47
Fair	+11·60	+9·00	−0·73	+28·00	+15·97	+0·87	+1·20	+9·27	+7·53	+12·47
Moderate	+11·50	+9·07	+2·13	+29·60	+18·50	+1·53	+1·00	+10·80	+10·00	+15·00
Good	+8·40	+10·80	+3·07	+28·47	+15·23	+3·40	+2·60	+12·53	+9·13	+14·80
Very good	+8·67	+15·00	+9·43	+25·07	+1·80	+3·07	+3·07	+11·60	+10·13	+12·33
Average excess for Set 2	+11·60	+10·92	+2·94	+26·99	+13·85	+1·75	+2·35	+10·55	+9·63	+12·81

Average of Average Excesses, 10·34

293. The Table shows clearly that, while in practically every case the examiners give higher marks when marking by details than when marking by impression, the higher marks are distributed unequally between the five groups.

Thus, Examiner A gives an average excess of 11·60 marks when marking by details over the marks obtained by impression ; this itself is an average of an excess of 17·83 marks for the Poor Group, 11·60 for the Fair Group, 11·50 for the Moderate Group, 8·40 for the Good Group, and 8·67 for the Very Good Group. In other words, in marking by details, he deals more generously with the Poor scripts than with the Good and Very Good scripts. But this is not true of all the examiners ; they vary in this matter in many ways.

294. The variations will be brought out most clearly by subtracting from each figure of the successive columns of Table 93 (corresponding to the five Groups) the average given at the bottom of the relevant column. We thus obtain :—

TABLE 94

AVERAGE DIFFERENCES BETWEEN DETAILED AND IMPRESSION MARKS GROUP BY GROUP WITH THE INFLUENCE OF DIFFERENT STANDARDS REMOVED

	Examiner A	B	C	E	G	K	L	M	N	P	Average
Poor	+6·23	−0·19	−2·14	−3·19	+3·88	−1·88	+1·52	−2·02	+1·70	−3·34	+0·06
Fair	0	−1·92	−3·67	+1·01	+2·12	−0·88	−1·15	−1·28	−2·10	−0·34	−0·82
Moderate	−0·10	−1·85	−0·81	+2·61	+4·65	−0·22	−1·35	+0·25	+0·37	+2·19	+0·57
Good	−3·20	−0·12	+0·13	+1·48	+1·38	+1·65	+0·25	+1·98	−0·50	+1·99	+0·50
Very Good	−2·93	+4·08	+6·49	−1·92	−12·05	+1·32	+0·72	+1·05	+0·50	−0·48	−0·32

295. It will be seen that Examiners C and K act in an opposite way to A. When marking by details they favour the Good

and Very Good scripts more than the Poor or Fair ones. Other examiners, e.g. L and M, distribute the additional marks which they give when marking by details fairly uniformly over all Groups. G, on the other hand, when marking by details adds on marks to all the Groups but the Very Good ; while E penalises the Poor and Very Good, but treats the middle Groups more favourably.

296. Let us now assume that the average differences of the marks allotted by X, Y and Z, as shown in Table 92, indicate the real differences in quality between corresponding Groups of Set 1 and Set 2 ; and let us compare these with the average differences for these Groups, assigned to them by Examiners A to P. The difference between these differences will show how far Examiners A to P as a whole have treated the different Groups, Poor to Very Good, differently in assigning to them the surplus of marks which they give when marking by details over the marks which they assign when marking by impression.

TABLE 95

	Averages from Table 94	Averages from Table 92	Difference
Poor	+0·06	+0·49	−0·43
Fair	−0·82	−0·51	−0·31
Moderate	+0·57	−0·71	+1·28
Good	+0·50	−0·29	+0·79
Very Good	−0·32	+1·02	−1·34

297. The final results show that in distributing the excess of marks which they allotted with the detailed method the examiners A to P on the whole tended to favour the scripts that were of medium and just above medium quality, and to undermark the other categories, especially the Very Good ; but, on the average, the marks involved do not amount to more than 1%.

APPENDIX TO CHAPTER VII

SPECIAL PLACE EXAMINATION (II)

ENGLISH ESSAY

I

Instructions to Examiners

1. Scripts 1–75 are to be marked *by impression only*. It is of the essence of the investigation that, in marking these scripts, no attempt should be made by the examiner to conform to the scheme of marking set out under II below or to any scheme of the kind. Examiners *are particularly requested to mark scripts* 1–75 *before they mark scripts* 76–150.

2. Scripts 76–150 are to be marked according to the amended marking-scheme set out under II below.

3. The maximum mark for all scripts is 100.

4. In order to facilitate the work, typed copies of all the scripts will be furnished to the examiners as well as the originals, and examiners will be free to make any marks they please on the typed scripts only. The originals are supplied for the allotment of marks in handwriting.

II

Amended Marking-Scheme

5. Marks are to be allotted as follows :—

(i)	Quantity, Quality, and Control of Ideas ...	50 marks
(ii)	Vocabulary	15 „
(iii)	Grammar and Punctuation	15 „
(iv)	Structure of Sentences	10 „
(v)	Spelling	5 „
(vi)	Handwriting	5 „
	Total	100 „

6. (i) *Quantity, Quality, and Control of Ideas*

It will be seen that half the total number of marks for the whole composition are to be given for the treatment of the

subject from a common-sense point of view. Has the candidate dealt sensibly with his subject ? Has he a fair number of ideas about it, and has he arranged them logically ? How far, in fine, has he shown evidence of *thought ?* These are the questions which have to be considered under the first head.

7. The mere length of the essay indicates roughly the quantity, as distinct from the quality, of the ideas. It must be remembered, however, that it gives some slight indication of quality as well. Several investigators have found that between the length and the merit of children's essays written within the same time there is a distinct positive correlation. Dr. William Boyd (whose book " Measuring Devices in Composition, Spelling and Arithmetic " gives valuable hints on the marking of essays) found the coefficient of correlation to be as high as ·69. That is interesting as a fact, but not very useful as a method ; for exceptions are numerous and flagrant. The point to be remembered is that in composition exercises a child of eleven writes *on an average from* 150 *to* 200 *words* in half an hour—a boy rather less, a girl rather more. That fact should be taken into account, and, other things being equal, marks should be added for an essay longer than the average and deducted for one that is shorter. First mark for quality and control of ideas on the ABCDE plan (see Section 11 below). Then add or deduct for quantity.

8. Length should specially be considered in *deducting marks for faults.* A child in writing 200 words would, at the same level of merit, make double the number of mistakes he would make in writing 100. Hence if you fix a maximum and deduct for faults, always make your deduction in respect of (say) the first 150 words *only.*

9. By control of ideas is meant the ability to arrange them logically and effectively. The examiner should ask himself—Is there evidence of some sort of plan, either written out beforehand or formulated in the mind ? Is the composition skilfully finished off ?

10. If a letter be chosen by the candidate, 6 marks should be deducted from the 50 for certain omissions and defects : 1 for an omitted address, 1 for an omitted date, 2 for faulty salutation, and 2 for faulty conclusion.

11. To mark for quality and control of ideas is no easy matter. The following procedure is recommended : Read the essay through, disregarding errors in the mere mechanics of English, and place it in one of the five categories, A, B, C, D, E, with A representing the highest and C the average. If your sample of scripts were really representative of the achievement of the

whole child population of the given age, the percentages falling into the various categories would be something like this :—

6	per	cent.	would	fall	into	Category	A
25	,,	,,	,,	,,	,,	,,	B
38	,,	,,	,,	,,	,,	,,	C
25	,,	,,	,,	,,	,,	,,	D
6	,,	,,	,,	,,	,,	,,	E

The scores, so far as they went, would then be normally distributed. But the budget of scripts dealt with by each examiner is *not* necessarily a representative sample. It is far too small for that. Moreover it is not a random sample ; the stupider children having been omitted, it is a selected sample. Then again there is the effect of teaching. Good teaching tends to skew the curve of distribution to the right and bad teaching to the left.

12. Hence it would be unreasonable to urge examiners to fit their marking to the normal curve of distribution. Still it would tend to standardise the marking if each examiner would, when he finds his marks departing *widely* from the above scheme, carefully consider whether he has not taken too high or too low a standard. This principle holds good with special force where natural ability, as distinct from acquired knowledge, is the main factor measured. Therefore it should not be applied with the same rigidity to the factors under (ii), (iii), (iv), (v) and (vi) as to those under (i).

13. Having settled the category you should now assign the appropriate marks. The following scheme is suggested :—

Category	Central Mark	Range of Marks
A	45·5	41–50
B	35·5	31–40
C	25·5	21–30
D	15·5	11–20
E	5	0–10

If the examiner would, when assigning the category, further differentiate by means of + or — he would find the marking facilitated. B + would mean 38–40 marks, while B — would mean 31–33.

14. By some such scheme as is here outlined, each essay will be given as just a rating as the inherent difficulty of the task permits.

(ii) *Vocabulary*

15. By vocabulary is meant not only the capacity to use unusual words, but also the capacity to use ordinary words

fittingly. A good plan of marking would be to take 8 as the central point of marking and to add or deduct marks according as the candidate's command of words falls above or below the average.

(iii) *Grammar and Punctuation*

16. Give 15 marks if there are no serious faults in grammar or punctuation, and deduct two for every serious fault. Do not be severe in marking mistakes in punctuation, apart from the misplacing of the full stop. It need scarcely be said that the lowest mark should be 0. Negative quantities are not to be used.

(iv) *Structure of Sentences*

17. Give a high mark if the essay contains a fair number of well-constructed sentences, a low mark if most of the sentences are loose and rambling. The use of participles, of relative pronouns, and of subordinate clauses are points of merit.

(v) *Spelling*

18. Deduct one half-mark for each serious mistake (*in the first* 150 *words only*) made in the spelling of common words only. Mistakes made in the spelling of uncommon words, and mistakes which you regard as obviously due only to hasty handwriting should be ignored.

(vi) *Handwriting*

19. If the handwriting is quite legible let it pass. Take marks off, however, for manifest carelessness.

CHAPTER VIII

COLLEGE ENTRANCE SCHOLARSHIP EXAMINATION: MARKING OF ENGLISH ESSAY

298. *Character of the Examination Paper.*—The paper was set at an Entrance Scholarship examination for a group of colleges in a University, and gave a choice of four subjects, on one of which an essay was to be written. No further direction was given. The time allowed was three hours.

299. *Selection of Scripts.*—Fifty scripts were selected from a larger number, so as to include most of those written by candidates who won Scholarships or Exhibitions. They comprised the scripts of all the ten candidates (Nos. 1–10) who had selected the first subject; of the eight (Nos. 11–18) who had selected the second; of the eleven (Nos. 19–29) who had selected the third; and of twenty-one (Nos. 30–50) of those who had selected the fourth.

300. *Procedure.*—After the removal of all marks on the essays indicating their origin and previous marking, they were submitted in turn to five distinguished experts in English from the same University, who were experienced in examining essays of this kind. Four of them had not previously seen these particular scripts. One of the examiners may have done so in the previous twelve months.

The examiners were asked to assign numerical marks with a maximum of 100; and they were also asked to assign a class to each candidate in accordance with the following scheme:—

Class I	67 marks and over
Class II	50 marks to 66 marks
Class III	33 marks to 49 marks
Class IV	Under 33 marks

In this way, an examiner who preferred to mark first in classes could so arrange his numerical marks as to fit in with his classification, while an examiner who preferred to mark numerically could arrange his classification to fit his marks. The important point

is that the classes of the candidates were determined by the examiners themselves, and were not merely the result of translating numerical marks into classes by a clerk in an office.

301. The numerical marks and classes assigned by the five examiners are set out in Table 96 below :—

TABLE 96

MARKS AND CLASSES AWARDED

No. of Candidate	No. of Subject taken	Examiner A		Examiner B		Examiner C		Examiner D[1]		Examiner E		Range
		Mark	Class	Mark	Class	Mark	Class	Mark	Class	Mark	Class	
1	1	45	III	38	III	20	IV	55	II	20	IV	35
2	1	43	III	44	III	55	II	48	III	52	II	12
3	1	70	I	67	I	65	II	60	II	60	II	10
4	1	30	IV	32	IV	35	III	40	III	30	IV	10
5	1	50	II	36	III	65	II	48	III	62	II	29
6	1	57	II	48	III	40	III	42	III	54	II	17
7	1	51	II	49	III	45	III	48	III	56	II	11
8	1	52	II	70	I	70	I	63	II	68	I	18
9	1	48	III	30	IV	55	II	55	II	40	III	25
10	1	70	I	65	II	55	II	66	II	64	II	15
11	2	65	II	74	I	65	II	55	II	58	II	19
12	2	54	II	58	II	55	II	50	II	63	II	13
13	2	67	I	50	II	45	III	42	III	52	II	25
14	2	62	II	53	II	60	II	42	III	56	II	20
15	2	55	II	55	II	65	II	66	II	63	II	11
16	2	69	I	57	II	70	I	63	II	65	II	13
17	2	72	I	65	II	60	II	55	II	40	III	32
18	2	65	II	51	II	60	II	63	II	50	II	15
19	3	45	III	45	III	45	III	60	II	42	III	18
20	3	32	IV	18	IV	35	III	33	III	30	IV	17
21	3	49	III	53	II	45	III	63	II	67	I	22
22	3	58	II	72	I	65	II	72	I	79	I	21
23	3	42	III	35	III	60	II	58	II	47	III	25
24	3	44	III	32	IV	45	III	52	II	45	III	20
25	3	60	II	32	IV	65	II	50	II	68	I	36
26	3	68	I	51	II	70	I	66	II	52	II	19
27	3	53	II	59	II	65	II	58	II	45	III	20
28	3	48	III	49	III	45	III	50	II	43	III	7
29	3	55	II	65	II	65	II	45	III	48	III	20
30	4	35	III	45	III	60	II	30	IV	45	III	30
31	4	51	II	59	II	55	II	60	II	50	II	10
32	4	54	II	51	II	55	II	48	III	50	II	7
33	4	60	II	63	II	35	III	63	II	56	II	28
34	4	65	II	52	II	40	III	60	II	55	II	25
35	4	69	I	68	I	60	II	63	II	68	I	9
36	4	53	II	65	II	55	II	55	II	58	II	12
37	4	50	II	72	I	60	II	65	II	64	II	22
38	4	45	III	65	II	65	II	58	II	51	II	20

[1] Examiner D divided Class II into two, II(i) and II(ii)—but as this division was not used by other examiners, it has been omitted from this Table.

TABLE 96—*continued*

No. of Candidate	No. of Subject taken	Examiner A		Examiner B		Examiner C		Examiner D		Examiner E		Range
		Mark	Class	Mark	Class	Mark	Class	Mark	Class	Mark	Class	
39	4	50	II	33	III	65	II	53	II	48	III	32
40	4	40	III	44	III	70	I	75	I	50	II	35
41	4	47	III	73	I	60	II	63	II	50	II	26
42	4	36	III	60	II	65	II	60	II	40	III	29
43	4	56	II	62	II	55	II	48	III	60	II	14
44	4	55	II	67	I	45	III	50	II	50	II	22
45	4	57	II	60	II	80	I	58	II	49	III	31
46	4	36	III	49	III	40	III	48	III	33	III	16
47	4	32	IV	36	III	35	III	55	II	30	IV	25
48	4	48	III	60	II	55	II	50	II	45	III	15
49	4	55	II	63	II	60	II	58	II	56	II	8
50	4	20	IV	33	III	35	III	10	IV	5	IV	30
	Average	51·9		52·7		54·8		54·0		50·6		20·0

302. The Table presents many features of interest. We may consider first of all the question of classes. In the award of a scholarship, what matters most is the allotment of First Classes, and the discrepancy between the examiners in respect of standard for a First Class is apparent in a number of ways. The following Table shows the statistical distribution of classes by the various examiners :—

TABLE 97

CLASSES AWARDED BY THE VARIOUS EXAMINERS

Examiner	1st Classes	2nd Classes	3rd Classes	4th Classes
A	7	24	15	4
B	8	23	14	5
C	5	29	15	1
D[1]	2	34	12	2
E	5	26	14	5

Whereas D allots only two First Classes, B allots eight (though the average mark of D, 54·0, slightly exceeds that of B, 52·7). D marks what he considers as the inferior essays more generously than B.

303. The general degree of agreement in regard to classes may be expressed in the following statement :—

All five examiners agreed on the class of seven candidates.
Four examiners agreed and one differed on the class of twelve candidates.
Three examiners agreed on one class, and the other two examiners agreed on another class in the case of eighteen candidates.

[1] Examiner D classed 16 as II (i) and 17 as II (ii), and one as II.

Three examiners agreed on one class and the other two examiners differed as to class both with the first three examiners and with one another in the case of seven candidates.

Two examiners agreed on one class, two examiners agreed on another class, and the other examiner placed the candidate in another class in the case of six candidates.

Thus there was complete agreement about the classification of a candidate in only 7 cases out of 50.

304. Table 98 below shows the awards of all the examiners to the twenty-five candidates who were allotted *either* a First Class *or* a Fourth Class by any one of them :—

TABLE 98

AWARDS OF CLASSES

Candidate	Examiner				
	A	B	C	D	E
*1	3	3	4	2	4
3	1	1	2	2	2
4	4	4	3	3	4
8	2	1	1	2	1
*9	3	4	2	2	3
10	1	2	2	2	2
11	2	1	2	2	2
*13	1	2	3	3	2
16	1	2	1	2	2
*17	1	2	2	2	3
20	4	4	3	3	4
*21	3	2	3	2	1
22	2	1	2	1	1
*24	3	4	3	2	3
*25	2	4	2	2	1
26	1	2	1	2	2
*30	3	3	2	4	3
35	1	1	2	2	1
37	2	1	2	2	2
*40	3	3	1	1	2
*41	3	1	2	2	2
*44	2	1	3	2	2
*45	2	2	1	2	3
*47	4	3	3	2	4
50	4	3	3	4	4

The candidates whose numbers are marked with an asterisk were placed in three different classes by different examiners. Perhaps the most striking instance of discrepancy is that of Candidate No. 25, who is given a First by Examiner E, but only a Fourth by Examiner B, although B is more generous with Firsts than any other examiner.

305. As the candidates with First Classes in the essay would probably be in the running for a scholarship (though the examination in English of course included papers of another type), it is specially interesting to see which candidates were recommended for a First by the different examiners, as set out in Table 99 below.

TABLE 99

Examiner				
A	B	C	D	E
Candidates				
3	3	8	22	8
10	8	16	40	21
13	11	26		22
16	22	40		25
17	35	45		35
26	37			
35	41			
	44			

306. It will be seen that not a single candidate out of the seventeen was placed in the First Class by more than three of the five examiners. Candidates Nos. 8, 22 and 35 each received three votes; Candidates Nos. 3, 16, 26 and 40 each had two votes; and the other ten had only one vote each. Thus the consensus of opinion in the cases that really matter is extraordinarily small.

307. We turn now to the numerical marks, which naturally show the same large differences between the estimates of the different examiners. The range of marks (i.e. the difference between the highest and lowest mark awarded to an individual candidate) varies from 7 to 36 and the average range is 20·0.

308. The extreme cases are shown below :—

MARKS AWARDED

Candidate	Examiner					Range
	A	B	C	D	E	
25	60	32	65	50	68	36
1	45	38	20	55	20	35
40	40	44	70	75	50	35

309. The averages of the different examiners for the paper as a whole and also for the different essay-subjects are set out in Table 100 below :—

TABLE 100

AVERAGES OF THE MARKS AWARDED BY THE DIFFERENT EXAMINERS[1]

Examiner	All Subjects	Subject 1	Subject 2	Subject 3	Subject 4
A	51·9 (4)	51·6 (2)	63·6 (1)	50·4 (4)	48·3 (4)
B	52·7 (3)	47·9 (5)	57·9 (3)	46·5 (5)	56·2 (1)
C	54·8 (1)	50·5 (4)	60·0 (2)	55·0 (2)	54·8 (2)
D	54·0 (2)	52·5 (1)	54·5 (5)	55·2 (1)	53·8 (3)
E	50·6 (5)	50·6 (3)	55·9 (4)	51·5 (3)	48·2 (5)
Range of Averages	4·2	4·6	9·1	8·7	8·0

310. The Table shows interesting features. The averages of the examiners for the whole series of essays are closer than might have been expected; they differ only by 4·2 marks. Moreover the order does not correspond at all to the number of First Classes which they awarded. The averages for the essays as a whole give little indication of the idiosyncrasies of the examiners.

The examiners are more differentiated by the averages they give for the essays on different subjects. Thus Examiner A, whose average for the essays as a whole ranks fourth, has the highest average (63·6) for Subject 2, and almost the lowest (48·3) for Subject 4. E (though he gives five Firsts) has the lowest average in two out of the five columns.

311. It seems that individual examiners tend to mark higher in one subject than another, so that a candidate may be handicapped by an unlucky choice of subject. The reasons for marking high or low for a particular subject can only be conjectured. It might be surmised that when the subject is already familiar to an examiner, he will be inclined to fix a particularly high standard for it; but on the other hand an unfamiliar subject may be one for which he has a distaste, and for which he is disposed to mark low on that account.

It would need a more extensive set of figures than the present ones to yield any definite conclusions upon the influence of the choice of subjects on the marks allotted.

[1] The figures in parentheses indicate the order of the examiners arranged according to the magnitude of their averages.

CHAPTER IX

MARKING OF UNIVERSITY MATHEMATICAL HONOURS SCRIPTS

312. *Character of the Examination Paper.*—The paper was one of university degree standard; it contained twelve questions, four relating to differential equations and eight relating to analytical geometry of three dimensions. Candidates were informed that they might attempt any number of questions, but that full marks might be obtained "on about six questions." Three hours was allowed for the paper.

313. *Object of the Investigation and Procedure.*—The object of the investigation was to test the degree of consistency (a) of individual examiners, all experienced in the particular kind of examination, and (b) of pairs of examiners, similarly experienced, and acting conjointly but independently of the other pairs.

For this purpose the twenty-three scripts written in answer to the question-paper were marked independently by Examiners A, B, C, D, E, and F, all capable and experienced examiners. The scripts were then independently revised by the pairs of examiners A and B, C and D, E and F, each pair achieving a set of revised marks. There were thus produced six sets of original marks and three sets of revised marks which were communicated by the authorities concerned who had arranged for the correction of the scripts in accordance with the plan described above.

314. The nine sets of marks (i.e. the original six sets of independent marks of the six examiners A, B, C, D, E, F, and the three sets of revised marks produced by the couples A, B; C, D; and E, F, acting conjointly) are set out in Table 101 below, together with the average marks for each examiner, and for each pair of examiners, and the ranges of marks for each can-

didate produced by the markings of the single examiners and of the pairs of examiners.

TABLE 101

MARKS AWARDED (Maximum = 300)

Candidate	Examiner A	B	C	D	E	F	Range	A,B	C,D	E,F	Range
1	209	185	223	235	225	212	50	198	230	219	32
2	200	205	180	193	205	208	28	203	183	207	24
3	201	208	172	198	197	179	36	203	186	190	17
4	175	193	172	177	212	189	40	186	177	210	33
5	81	94	81	100	123	145	64	86	96	128	42
6	200	217	203	205	207	187	30	207	208	195	13
7	119	140	137	157	134	150	38	125	145	142	20
8	167	201	187	198	190	190	34	188	194	190	6
9	147	155	127	139	140	147	28	151	138	144	13
10	203	220	203	192	205	208	28	216	203	207	13
11	85	66	79	78	108	65	43	76	87	88	12
12	133	122	140	128	127	133	18	128	137	130	9
13	224	228	239	253	222	241	31	220	246	239	26
14	215	226	228	223	234	217	19	220	226	225	6
15	224	245	255	262	216	245	46	239	260	241	21
16	95	120	136	143	135	127	48	117	136	131	19
17	165	161	171	168	178	177	17	163	171	178	15
18	287	294	290	308	300	303	21	290	300	302	12
19	123	101	66	100	114	102	57	113	91	108	22
20	154	125	118	122	163	175	57	132	123	169	46
21	117	102	120	131	136	113	34	110	120	122	12
22	89	73	75	81	75	87	16	79	83	81	4
23	271	278	277	287	273	282	16	279	282	278	4
Average	168·9	172·1	168·7	177·3	179·1	177·5	34·7	170·8	174·9	179·3	18·3
Mean Deviations	48	55	53	52	47	46		52	52	48	

315. In the individual markings it will be seen that the lowest averages are those of Examiners A and C, which are almost identical, 168·9 and 168·7, and the highest average is that of E, 179·1. The difference of about 11 marks is just under 4% of the maximum.

The extreme difference of the averages of the three pairs of examiners (A, B), (C, D), (E, F) (varying from 170·8 to 179·3) is about 9 marks.

316. The mean deviations, measuring the spread of the marks, are roughly the same for each examiner, and for each pair of examiners. There is no evidence here that when pairs of examiners allot marks they necessarily award marks with a smaller spread than when they act individually.

317. But these differences of averages yield very little indication of the differences of the marks which were allotted to individual candidates. It will be seen that the six independent markings of Examiners A to F yield ranges of which the lowest is 16 and the highest 64, out of a maximum of 300, with an average of 34·7. One may ask fairly what is the validity of any one judgment when the judgments of the individual examiners vary so greatly.

318. The procedure of settling marks on the verdict of two examiners acting concurrently reduced the extreme difference of the averages only slightly; but it had a much greater effect in reducing the ranges, of which the extremes for the pairs are 4 and 46 and the average 18·3, only a little more than half the average range for the six examiners (34·7). But the fact that in an examination of this kind two out of three pairs of examiners can differ by as much as they do in the case of Candidate No. 20, who is assigned 132, 123 and 169 marks, or of Candidate No. 4, who is assigned 186, 177 and 210 marks, is remarkable.

319. The results of the present investigation are particularly interesting, because it is commonly supposed that mathematics papers are easily examined, and that not much error is likely to be introduced into the results by having the scripts marked by different examiners. The analysis shows, however, that, if justice is to be done, quite a considerable number of scripts should receive extra consideration.

When the candidates are placed in order of merit by the original six examiners, the results are even more illuminating than when the original marks themselves are considered.[1] All the examiners agree in the placing of the two candidates at the top of the group of twenty-three, and in placing the 13th in order of merit. The examiners do not agree in the placing of the other twenty. The pairs agree in the placing of five candidates, the 1st, 2nd, 3rd, 4th, and 13th, but disagree in the order of merit of the other eighteen.

[1] It should be stated that the examination was not one on which a scholarship was awarded, and that the examiners themselves did not therefore consider the order of merit.

The order in which the candidates are placed by the examiners is given below :—

TABLE 102

ORDER OF MERIT

No. of Candidate	Examiners A	B	C	D	E	F	Greatest difference of place	Pairs of Examiners AB	CD	EF	Greatest difference of place
1	6	12	6	5	4	6	8	10	5	6	5
2	9½	9	10	10	9½	7½	2½	8½	11	8½	2½
3	8	8	11½	8½	11	12	4	8½	10	11½	3
4	11	11	11½	12	7	10	5	12	12	7	5
5	23	21	20	20½	20	17	6	21	20	19	2
6	9½	7	7½	7	8	11	4	7	7	10	3
7	18	15	15	14	18	15	4	17	14	16	3
8	12	10	9	8½	12	9	3½	11	9	11½	2½
9	15	14	17	16	15	16	3	14	15	15	1
10	7	6	7½	11	9½	7½	5	6	8	8½	2½
11	22	23	21	23	22	23	2	23	22	22	1
12	16	17	14	18	19	18	5	16	16	18	2
13	3½	4	4	4	5	4	1½	4½	4	4	0½
14	5	5	5	6	3	5	3	4½	6	5	1½
15	3½	3	3	3	6	3	3	3	3	3	0
16	20	18	16	15	17	19	5	18	17	17	1
17	13	13	13	13	13	13	0	13	13	13	0
18	1	1	1	1	1	1	0	1	1	1	0
19	17	20	23	20½	21	21	6	19	21	21	2
20	14	16	19	19	14	14	5	15	18	14	4
21	19	19	18	17	16	20	4	20	19	20	1
22	21	22	22	22	23	22	2	22	23	23	1
23	2	2	2	2	2	2	0	2	2	2	0

320. It is clear that the pairing of the examiners notably diminishes the differences of the order in which the candidates are placed. But it is interesting to note what considerable differences may still subsist. Thus Candidate No. 1, whose place varies with the individual examiners from 4th (Examiner E) to 12th (Examiner B) of the twenty-three candidates, is placed 10th by the pair AB (marks 198), 5th by the pair CD (marks 230) and 6th by the pair EF (marks 219), whereas Candidate No. 4 is placed 12th by the pair AB (marks 186), 12th by the pair CD (marks 177) and 7th (marks 210) by the pair EF. The pair of examiners AB and the pair EF regard Candidate No. 1 and Candidate No. 4 as not being very different in merit, compared to each other, though they put them in very different places among their co-examinees ; while the pair of examiners CD regard them as differing widely.

CHAPTER X

MARKING OF UNIVERSITY HISTORY HONOURS SCRIPTS

321. *Character of the Examination Papers.*—The examination papers were four in number, all forming part of a University History Honours Examination. The subjects of the papers were as follows :—

Paper I. Ancient and Mediæval History.

Paper II. Mediæval and Modern History.

Paper III. An Essay paper with a choice from a number of subjects.

Paper IV. Political Thought (Prescribed Books).

In Papers I, II and IV, candidates were requested not to attempt more than four questions out of a considerable number. The time allowed for each paper was three hours.

322. *Procedure.*—The University concerned furnished us with all the scripts available in the subjects enumerated above from a recent Honours examination.[1] Unfortunately three scripts (which happened to be among the best) had been accidentally destroyed. The total number of scripts available was eighteen for Paper I, seventeen for Paper II, eighteen for Paper III, and sixteen for Paper IV.

The following seventeen examiners took part in the marking of the scripts :—

Professor J. B. Black, M.A., Burnett-Fletcher Professor of History in the University of Aberdeen.

[1] The examination included a number of other papers, but it was thought that the field covered by these was sufficient for the purpose of the investigation.

PROFESSOR A. BROWNING, M.A., D.Lit., Professor of History in the University of Glasgow.

MR. NOEL DENHOLM-YOUNG, M.A., Fellow of Magdalen College, Oxford.

PROFESSOR A. H. DODD, M.A., Professor of History in the University of Wales.

MR. D. L. KEIR, M.A., Fellow of University College and University Lecturer in English Constitutional History, Oxford.

MR. R. B. MCCALLUM, M.A., Fellow and Lecturer in Modern History, Pembroke College, Oxford.

PROFESSOR J. L. MORISON, M.A., D.Lit., Professor of Modern History, Armstrong College, University of Durham.

PROFESSOR R. B. MOWAT, M.A., Professor of History in the University of Bristol.

MR. J. N. L. MYRES, M.A., Student and Tutor of Christ Church, Oxford.

MR. E. J. PASSANT, M.A., Fellow of Sidney Sussex College, Cambridge.

MISS I. G. POWELL, M.A., Lecturer in History at the Royal Holloway College, University of London.

PROFESSOR EILEEN POWER, M.A., D.Lit., Professor of Economic History in the University of London.

PROFESSOR F. M. POWICKE, Litt.D., F.B.A., Regius Professor of Modern History in the University of Oxford.

MR. G. H. STEVENSON, M.A., Fellow of University College and University Lecturer in Ancient History, Oxford.

MR. C. G. STONE, M.A., Balliol College, Oxford.

PROFESSOR A. F. BASIL WILLIAMS, O.B.E., M.A., F.B.A., Professor of History in the University of Edinburgh.

PROFESSOR C. H. WILLIAMS, M.A., Professor of History in the University of London.

The examiners are designated A, B, C, . . . R, in what follows, but this designation does not correspond with the alphabetical order of the names.

323. The scripts of Paper I were marked by five examiners; the scripts of each of the other papers by ten examiners. The only reason for having the scripts of Paper I marked by fewer examiners was the difficulty in getting examiners to cover the two periods with which it dealt.

As in other investigations, no indication of origin or of the original marking appeared on the scripts, or was communicated to the examiners.

Each examiner marked each individual question separately and gave a final mark for each script as a whole.

324. The following "literal" system of marking, including 24 grades, ranging from δ to α+, was, after consultation with an eminent historian, submitted to and approved by the great majority of examiners before the work began. It was communicated as approved to one or two examiners who came into the investigation subsequently.

TABLE 103

Literal Mark	No. of Grade	Literal Mark	No. of Grade	Literal Mark	No. of Grade
α+	(24)	β++	(15)	βγ	(6)
α?+	(23)	β+?+	(14)	γβ	(5)
α	(22)	β+	(13)	γ+	(4)
α?−	(21)	β?+	(12)	γ	(3)
α−	(20)	β	(11)	γ−	(2)
α−?−	(19)	β?−	(10)	δ	(1)
α=	(18)	β−	(9)		
αβ	(17)	β−?−	(8)		
βα	(16)	β=	(7)		

325. It may be well to say a word here on the use of a literal system of this kind as compared with the numerical systems employed in our other investigations. The literal system is generally used at Oxford; there is a considerable variety of usage in other Universities.

326. There seems to be a fundamental difference, at any rate at the first blush, between the two systems. The literal system indicates only an order in classification, not ratios of proficiency. With that system, there can be no question of adding up marks for individual questions in order to obtain a percentage of a total maximum. It would appear that the literal mark indicates in the examiner's mind a certain "quality." The question of "quantity" probably enters into his estimate only in a subordinate degree.

With the numerical system, on the other hand, the marks for individual questions are added up to furnish a total, a procedure which is convenient, though it is based on hypotheses which it is not perhaps easy to analyse and justify. But any attempt to add together the symbols indicating "classes" or "grades" would seem *a priori* unjustifiable and would be rejected by many who use literal marks.

327. Both systems have their conveniences. It is for the sake of readers who are unaccustomed to literal marking, and to enable them to estimate by what number of grades (or subordinate "classes") any two examiners differ, that we have attributed

he numbers 1 to 24 to the successive grades, δ to α+, and that, ide by side with the literal tables, we have inserted numerical ables on this basis. But, for the reasons stated above, the umbers indicating grades must not be regarded as numerical narks. They are ordinal numbers, not cardinal.

328. Readers accustomed to numerical marking may further wish to have some means of comparison between the two systems. A rough and ready form of translation from one into the other would be to suppose that each of the 24 literal symbols corresponds to a multiple of four marks, and the highest, α+, to 96. Only an experimental investigation could afford any real basis for such a translation. But it is certain that such a difference as that of 18 grades, the maximum difference between the awards of two different examiners to the same script in this investigation, much more nearly approaches a difference of 72 in numerical marking, with 96 (or 100) as a maximum mark, than a difference of 18, which a superficial glance might suggest.

329. An index of the examiners who marked the various papers is given in the Table below :—

TABLE 104

Examiner	Paper I	II	III	IV
A	–	*	*	*
B	–	*	*	*
C	–	*	–	–
D	*	–	–	–
E	–	–	–	*
F	–	*	*	*
G	–	–	–	*
H	–	*	*	*
J	–	*	*	*
K	*	*	*	–
L	–	*	*	*
M	–	–	–	*
N	–	*	*	–
O	*	–	–	–
P	*	–	–	–
Q	*	–	*	*
R	–	*	*	–

The papers marked by each examiner are indicated by an asterisk in the row corresponding to the letter by which he is designated. Thus Examiner B marked Papers II, III and IV.

330. In Tables 105, 106, 106A, 107, 107A, 108, and 108A are set out the literal marks assigned by the examiners to the scripts of each candidate, and the numerical representation of the corresponding grades according to the convention explained in paras. 327-328 above.

TABLE 105

PAPER I

Examiner		*Marks allotted*					*Numerical representation of the marks in ordered grades*						
		D	K	O	P	Q	D	K	O	P	Q	*Range in grades*	*Range in grades neglecting Q's results*
Cand. No.	1	β++	β?+	β	β=	γβ	15	12	11	7	5	10	8
	2	αβ	β?−	β+	β?+	β	17	10	13	12	11	7	7
	3	β?−	β+	β?−	β−	γβ	10	13	10	9	5	8	4
	4	β−	β++	β	β?−	γ	9	15	11	10	3	12	6
	5	β+	β?−	β−	γ+	γ+	13	10	9	4	4	9	9
	6	αβ	β+?+	α−	β+	γβ	17	14	20	13	5	15	7
	7	β−	β?+	β?+	γβ	βγ	9	12	12	5	6	7	7
	8	β?+	β+	β	γ+	γ+	12	13	11	4	4	9	9
	9	α−	β	β−	β+	β+	20	11	9	13	13	11	11
	10	β−	β+	β	β−	β?+	9	13	11	9	12	4	4
	11	γ	γ	γ−	γ−	γ−	3	3	2	2	2	1	1
	12	γβ	β+?+	β	β−	γβ	5	14	11	9	5	9	9
	13	αβ	αβ	α	γβ	βγ	17	17	22	5	6	17	17
	14	βα	β++	β++	β+	β?−	16	15	15	13	10	6	3
	15	β−	γ+	β	γ	γ	9	4	11	3	3	8	8
	16	β	β+?+	β	γ	β+	11	14	11	3	13	11	11
	17	β=	β	β	β+	γβ	7	11	11	13	5	8	6
	18	α	β++	α?−	β	β	22	15	21	11	11	11	11
Median		β / β?+	β+	β	β−	γβ	11–12	13	11	9	5	Average 9·1	Average 7·7

TABLE 106

PAPER II

Marks allotted

Examiner		A	B	C	F	H	J	K	L	N	R
Cand. No.	1	β+?+	β−	α−?−	β−	βα	β	β	βγ	γβ	β?−
	2	β?+	β+	βα	αβ	β++	β+	βα	β−	β+?+	β−
	3	β	β+?+	β	β−	β+	β−	β++	β=	βγ	βα
	4	β−	α=	αβ	β+?+	β+	γ+	β+?+	β?−	β	β+
	5	β++	β++	β?−	β?+	β?+	γ+	β?−	β?−	γβ	β?+
	6	β++	α−?−	β?+	β+?+	β+?+	β?+	β−	β+	β	β++
	7	βα	αβ	β+	β?−	βγ	β−	γ	β?+	β	β?+
	8	β	α−	β	β?−	β−	γ+	β+	β−	β+?+	β++
	9	α=	αβ	β−	βγ	βγ	β	β+	β?−	β+	αβ
	10	β+	β+	β+?+	βα	γβ	β−	β+?+	β=	γβ	β++
	11	δ	β=	β−	β−?−	γ+	β−	βγ	βγ	δ	γβ
	12	β?+	β?+	αβ	β	γ+	β	β+?+	β?−	β++	αβ
	13	β++	β+	β?−	β++	β?+	γβ	β++	β−	β?+	β+?+
	14	β+?+	α−	α=	βα	γβ	β++	βα	β+	βα	β+
	15	β?−	γ+	β?−	β−	γ?+	β=	γ+	β?+	β=	β?−
	17	β?+	β+	β	β+	β?−	β=	β?+	β=	β	β?+
	18	β++	β++	β?+	βα	β?+	β+	β+	β+?+	βα	α−
Median		β+	β+?+	β?+	β?+	β?−	β−	β+	β?−	β	β+

TABLE 106A
PAPER II

Numerical representation of the marks in ordered grades.

Examiner		A	B	C	F	H	J	K	L	N	R	*Range in grades*
Cand. No.	1	14	9	19	9	16	11	11	6	5	10	14
	2	12	13	16	17	15	13	16	9	14	9	8
	3	11	14	11	9	13	9	15	7	6	16	10
	4	9	18	17	14	13	4	14	10	11	13	14
	5	15	15	10	12	12	4	10	10	5	12	11
	6	15	19	12	14	14	12	9	13	11	15	10
	7	16	17	13	10	6	9	3	12	11	12	14
	8	11	20	11	10	9	4	13	9	14	15	16
	9	18	17	9	6	6	11	13	10	13	17	12
	10	13	13	14	16	5	9	14	7	5	15	11
	11	1	7	9	8	4	9	6	6	1	5	8
	12	12	12	17	11	4	11	14	10	15	17	13
	13	15	13	10	15	12	5	15	9	12	14	10
	14	14	20	18	16	5	15	16	13	16	13	15
	15	10	4	10	9	3½	7	4	12	7	10	8½
	17	12	13	11	13	10	7	12	7	11	12	6
	18	15	15	12	16	12	13	13	14	16	20	8
Median		13	14	12	12	10	9	13	10	11	13	Average 11·1

TABLE 107
PAPER III

Marks allotted.

Examiner		A	B	F	H	J	K	L	N	Q	R
Cand. No.	1	$\beta\alpha$	$\beta?+$	β	$\alpha=$	$\beta+$	$\beta+$	$\beta?+$	β	$\gamma\beta$	$\beta+$
	2	$\beta-$	$\beta+$	$\alpha=$	$\beta+$	$\beta?+$	β	$\beta?+$	$\beta-$	$\beta+?+$	$\beta+$
	3	$\beta?+$	$\beta++$	β	$\beta++$	β	$\beta+?+$	$\beta+$	$\beta\gamma$	$\beta\gamma$	α
	4	$\beta?-$	β	β	β	$\beta?+$	β	$\beta?-$	$\beta\gamma$	γ	$\beta?+$
	5	$\beta\alpha$	$\gamma+$	$\beta-$	$\beta-$	$\beta+$	$\beta+$	$\beta?-$	$\gamma\beta$	$\gamma+$	$\beta-$
	6	$\beta++$	$\alpha-$	$\beta\alpha$	$\beta+$	$\beta++$	$\beta?+$	$\beta+?+$	$\beta?+$	$\gamma\beta$	$\alpha\beta$
	7	β	β	$\beta-$	β	$\beta-$	$\beta=$	β	$\beta=$	$\beta\gamma$	$\beta+$
	8	β	$\alpha=$	β	β	$\beta+$	$\beta-$	$\beta+$	$\beta=$	$\beta\gamma$	$\beta+$
	9	α	$\gamma+$	$\beta+$	$\gamma\beta$	$\alpha-$	$\beta+?+$	$\beta+$	$\beta+$	$\alpha\beta$	$\alpha-$
	10	$\beta?+$	$\beta+$	$\beta++$	$\beta-$	$\gamma+$	$\beta++$	$\beta?-$	β	$\beta++$	$\beta?+$
	11	δ	γ	$\beta=$	β	$\gamma-$	$\beta\gamma$	$\gamma\beta$	δ	$\gamma-$	$\gamma+$
	12	$\beta?+$	$\gamma+$	$\beta+$	β	β	$\beta?-$	$\beta\gamma$	$\beta=$	$\gamma+$	β
	13	$\beta+$	$\beta-$	$\alpha-$	$\beta+$	$\beta+$	$\beta\alpha$	$\beta?-$	$\beta+$	$\beta-$	$\beta?+$
	14	$\beta+?+$	$\beta\alpha$	$\alpha\beta$	$\gamma+$	$\gamma+$	$\beta+?+$	$\beta?+$	β	$\beta+$	$\beta\alpha$
	15	β	$\beta++$	$\beta-$	$\gamma\beta$	β	$\gamma+$	$\beta=$	β	$\gamma+$	$\beta+?+$
	16	$\beta+$	$\alpha-$	$\alpha\beta$	$\beta\alpha$	$\beta\alpha$	$\beta?+$	$\beta\alpha$	$\beta+?+$	$\alpha\beta$	$\beta\alpha-$
	17	$\beta?+$	$\gamma+$	$\beta=$	$\beta?+$	γ	γ	$\beta\gamma$	γ	γ	$\beta?-$
	18	β	$\beta++$	$\alpha=$	$\beta++$	$\beta++$	β	$\beta\alpha$	$\beta+$	$\beta\alpha$	$\alpha?+$
Median		$\beta?+$	$\beta+$ / $\beta?+$	$\beta?+$	β	$\beta?+$	β / $\beta?+$	β / $\beta?+$	$\beta?-$	$\beta\gamma$	$\beta+$

TABLE 107A
PAPER III

Numerical representation of the marks in ordered grades

Examiner		A	B	F	H	J	K	L	N	Q	R	*Range in grades*	*Range in grades neglecting Q's results*
Cand. No.	1	16	12	11	18	13	13	12	11	5	13	13	7
	2	9	13	18	13	12	11	12	9	14	13	9	9
	3	12	15	11	15	11	14	13	6	6	22	16	16
	4	10	11	11	11	12	11	10	6	3	12	9	6
	5	16	4	9	9	13	13	10	5	4	9	12	12
	6	15	20	16	13	15	12	14	12	5	17	15	8
	7	11	11	9	11	9	7	11	7	6	13	7	6
	8	11	18	11	11	13	9	13	7	6	13	12	11
	9	22	4	13	5	20	14	13	13	17	20	18	18
	10	12	13	15	9	4	15	10	11	15	12	11	11
	11	1	3	7	11	2	6	5	1	2	4	10	10
	12	12	4	13	11	11	10	6	7	4	11	9	9
	13	13	9	20	13	13	16	10	13	9	12	11	11
	14	14	16	17	4	4	14	12	11	13	16	13	13
	15	11	15	9	5	11	4	7	11	4	14	11	11
	16	13	20	17	16	16	12	16	14	17	$15\frac{1}{2}$	8	8
	17	12	4	7	12	3	3	6	3	3	10	9	9
	18	11	15	18	15	15	11	16	13	16	23	12	12
Median		12	12–13	12	11	12	11–12	11–12	10	6	13	Average 11·4	Average 10·4

TABLE 108

PAPER IV

Marks allotted

Examiner		A	B	E	F	G	H	J	L	M	Q
Cand No.	1	β	$\beta?+$	$\beta-$	$\beta?-$	$\beta?-$	$\beta?+$	$\beta+?+$	$\beta?-$	$\beta+?+$	$\beta\gamma$
	2	$\beta?-$	$\beta-$	$\beta?+$	$\beta\alpha$	$\beta+$	$\beta+$	$\alpha=$	$\beta?+$	$\beta+$	$\beta?+$
	3	$\beta+$	$\beta\div$	$\beta\div$	β	$\beta+?+$	$\beta-$	$\beta++$	$\beta?+$	$\beta?-$	$\beta-$
	4	$\beta+?+$	β	$\beta\alpha$	$\beta?+$	$\beta-?-$	β	$\beta+$	$\beta\gamma$	$\beta+$	$\beta\gamma$
	5	β	$\beta?-$	$\beta?-$	$\beta=$	$\beta-$	$\gamma\div$	$\gamma+$	$\gamma\beta$	$\gamma\beta$	$\gamma\div$
	6	$\beta\alpha$	$\beta+$	$\beta?-$	$\beta?-$	$\alpha=$	$\beta\gamma$	$\beta++$	$\beta?+$	$\beta+?+$	$\gamma\beta$
	7	$\beta\gamma$	$\gamma\div$	$\beta=$	$\beta?-$	$\beta?+$	$\beta\gamma$	$\gamma+$	$\beta=$	β	γ
	8	$\beta\gamma$	$\alpha-$	$\beta+$	$\beta+$	$\beta=$	$\beta?+$	$\gamma+$	$\beta?+$	$\beta+$	$\beta\gamma$
	9	α	$\beta\div$	$\beta\alpha$	$\beta?-$	$\alpha\beta$	$\beta?+$	$\beta+$	$\beta+$	$\beta\alpha$	$\beta+?+$
	10	$\beta\alpha$	$\beta?\div$	$\beta?\div$	$\alpha\beta$	$\beta?-$	$\beta+$	$\beta?+$	$\beta-$	$\beta+$	$\beta+?+$
	11	δ	γ	$\gamma+$	$\beta?-$	$\beta-$	γ	$\gamma+$	$\gamma\beta$	$\beta=$	$\beta\gamma$
	12	$\beta+?+$	β	$\beta?+$	$\beta?+$	$\beta?\div$	$\beta\gamma$	$\beta+$	$\beta\gamma$	β	$\gamma+$
	13	$\beta\alpha$	$\alpha-$	$\alpha-?-$	$\alpha=$	$\beta?\div$	$\beta+\div$	$\beta+?+$	$\beta=$	$\alpha\beta$	$\beta+$
	15	$\beta\div?+$	$\beta\div$	$\alpha\beta$	$\beta?+$	$\beta-$	$\beta+$	$\gamma+$	$\beta-$	$\beta++$	$\beta-$
	17	β	β	$\beta-$	$\beta-?-$	$\beta-$	$\beta=$	$\beta+$	$\beta=$	$\beta+$	$\gamma+$
	18	$\beta?-$	$\beta++$	$\alpha\beta$	$\beta\alpha$	$\alpha\beta$	$\alpha-$	$\beta?+$	$\beta+?+$	$\beta+?+$	$\alpha\beta$
Median		$\beta?\div$	$\beta?+$	$\beta?\div$	β / $\beta?+$	$\beta?+$	β / $\beta?+$	$\beta+$	$\beta-$	$\beta+$	$\beta\gamma$

TABLE 108A

PAPER IV

Numerical representation of the marks in ordered grades

Examiner	A	B	E	F	G	H	J	L	M	Q	*Range in grades*	*Range in grades neglecting Q's results*
Cand. No. 1	11	12	9	10	10	12	14	10	14	6	8	5
2	10	9	12	16	13	13	18	12	13	12	9	9
3	13	13	13	11	14	9	15	12	10	9	6	6
4	14	11	16	12	8	11	13	6	13	6	10	10
5	11	10	10	7	9	4	4	5	5	4	7	7
6	16	13	10	10	18	6	15	12	14	5	13	12
7	6	4	7	10	12	6	4	7	11	3	9	8
8	6	20	13	13	7	12	4	12	13	6	16	16
9	22	13	16	10	17	12	13	13	16	14	12	12
10	16	12	12	17	10	13	12	9	13	14	8	8
11	1	3	4	10	9	3	4	5	7	6	9	9
12	14	11	12	12	12	6	13	6	11	4	10	8
13	16	20	19	18	12	15	14	7	17	13	13	13
15	14	13	17	12	13	13	4	9	15	9	13	13
17	11	11	9	8	13	7	13	7	13	4	9	6
18	10	15	17	16	17	20	12	14	14	17	10	10
Median	12	12	12	11–12	12	11–12	13	9	13	6	Average 10·1	Average 9·5

331. A glance at the Tables shows certain general features of interest. We have a closeness of marking between certain examiners and a wide difference between others, not attributable to chance, but showing real and probably irreconcilable differences of standard.

332. The examiners were asked to indicate what were their limits for a First, a Second, and a Third Class. Not all replied on the point. In the original scheme, a copy of which was furnished to each examiner (see para. 324), there was a gap between $\beta\alpha$ and $\beta++$, and between $\beta=$ and $\beta\gamma$, there being tacitly implied three classes. The following is a summary of information supplied by the examiners on the meaning of the symbols.

A :—$\alpha\beta$ and $\beta\alpha$ borderline. So also $\beta\gamma$ and $\gamma\beta$. δ fails.

B :—Nil.

C :—$\alpha\beta$ is a first, $\beta\alpha$ a second, $\beta\gamma$ is a third class. δ is a failure. Rarely uses high α's, or low marks, e.g. γ's.

D :—Does not use $\alpha+$ or $\alpha?+$, perfection is α. $\beta\alpha$ or $\beta++$ is the best second class. He would have put $\beta\alpha$ at the top of the second group.

E :—$\alpha\beta$ and $\beta\alpha$ are borderline marks, the former indicating a first class paper with either one poor answer or one persistent fault, the other a second class paper with one excellent answer or one very sound quality. Similarly with other borderline marks. Failures are $\gamma-$ and δ.

F :—$\beta\alpha$ is top of second class. $\beta?-$ is top of third. δ is failure.

G :—$\beta\alpha$ is top of second, $\beta=$ is top of third class, $\alpha\beta$ and $\beta\alpha$ are borderline and $\beta-?-$ is borderline. $\gamma-$ and δ are failures.

H :—$\alpha\beta$ and $\beta\alpha$ as in E. δ is failure.

J :—First, second and third class as implied in the scheme sent out.

K :—Nil.

L :—$\alpha\beta$ minimum for first class. $\beta\alpha$ borderline. $\beta\gamma$ minimum for second. $\gamma\beta$ borderline. δ failure.

M :—$\alpha\beta$ minimum for first class. $\beta\alpha$ borderline. $\beta\gamma$ and $\gamma\beta$ borderline. δ failure.

N :—$\alpha\beta$ minimum for first class, $\beta\alpha$, second. Third, $\beta\gamma$ to and including δ.

O :—As in E with qualification "that value of borderline marks as means of judging is that, if several papers have to be assessed in the final result, the mixed or 'border' marks have an additional significance, pointing to the need for inquiry. They *suggest* quality. Hence I should personally avoid them if only one paper was set on a subject."

P :—$\alpha\beta$ and $\beta\alpha$ borderline as E. So with the $\beta\gamma$ and $\gamma\beta$.

Q :—$\alpha\beta$ minimum for first class. $\beta\alpha$ highest second. So with others.

R :—$\alpha\beta$ and $\beta\alpha$ borderline. $\beta-$, $\beta-?-$, $\beta=$ borderline. $\beta\gamma$ highest third class. $\gamma-$ and δ fail.

333. The examiners are not in sufficient agreement on this point to use their remarks as a basis for classification. In actual practice it is well known that the limits are not determined in any purely mechanical way, but are the subjects of discussion in connexion with all borderline cases. The subject of the present investigation is not the actual award of First, Second and Third

Classes at a History Honours examination, but the variation in the individual judgments which must serve as a basis for those awards.

Although we cannot use the terms First, Second and Third Class, we can distinguish between the number of α's, β's, γ's, and δ's and of borderlines.

Thus the lowest limit for a First Class most generally adopted is $\alpha\beta$; but some are willing to consider $\beta\alpha$, the next grade, as a borderline for a First.

There is much more variation in the opinions as to the lower limit of a Second Class :—

β is adopted by F,
$\beta=$, by C, H, J, and N,
$\beta\gamma$, by Q.

Some of the other examiners indicate that the borderline marks between second and third class are as follows :—

$\beta-$, $\beta-$?$-$, $\beta=$, Examiner R
$\beta-$?$-$, Examiner G
$\beta\gamma$ and $\gamma\beta$, Examiners A, E, M, P
$\gamma\beta$, Examiner L.

We have thus a difference of several grades between the highest and the lowest limit adopted by the different examiners. In the Tables below we treat as α's the grades from $\alpha+$ to $\alpha=$, as β's the grades from $\beta++$ to $\beta=$; as γ's the grades from $\gamma+$ to $\gamma-$. $\alpha\beta$ and $\beta\alpha$ are treated as borderline cases between α and β ; and $\beta\gamma$ and $\gamma\beta$ as borderline cases between β and γ.

334. We give in Tables 109 to 112 below the classification statistics of the various examiners on the foregoing basis, for the scripts marked by them.

TABLE 109

PAPER I (Ancient and Mediæval History)

Marks	Examiner D	K	O	P	Q
			Number of Awards		
α	2		3		
Borderline	4	1			
β	10	15	14	11	6
Borderline	1			2	7
γ	1	2	1	5	5
δ					
	18	18	18	18	18
Median	β / β?+ (11-12)	$\beta+$ (13)	β (11)	$\beta-$ (9)	$\gamma\beta$ (5)

Thus Examiner D gives two candidates clear α's, four candidates a borderline mark between α and β, ten candidates β, one candidate $\gamma\beta$, and one candidate γ. Q returns them all as β or worse, and no examiner uses δ.

335. TABLE 110

PAPER II (Mediæval and Modern History)

Marks	Examiner									
	A	B	C	F	H	J	K	L	N	R
	Number of Awards									
α	1	4	2							1
Borderline	1	2	3	4	1		2		2	3
β	14	10	12	12	9	13	12	15	10	12
Borderline				1	4	1	1	2	4	1
γ		1			3	3	2			
δ	1								1	
	17	17	17	17	17	17	17	17	17	17
Median	$\beta+$ (13)	$\beta+$ (14)	?+ (12)	β?+ (12)	β?+ (10)	β?− (9)	$\beta-$ (13)	$\beta+$ (10)	β?− (11)	β (13)

J and L mark the scripts as β or worse, C as β or better. A and N are the only ones to use δ.

TABLE 111

336. PAPER III (Essay)

Marks	Examiner									
	A	B	F	H	J	K	L	N	Q	R
	Number of Awards									
α	1	3	3	1	1					3
Borderline	2	1	3	1	1	1	2		3	3
β	14	9	12	13	12	14	13	13	4	11
Borderline				2		1	3	3	5	
γ		5		1	4	2		1	6	1
δ	1							1		
	18	18	18	18	18	18	18	18	18	18
Median	β?+ (12)	$\beta+$ / β?+ (12-13)	β?+ (12)	β (11)	β?+ (12)	β / β?+ (11-12)	β / β?+ (11-12)	β?− (10)	$\beta\gamma$ (6)	$\beta+$ (13)

N marks all the candidates as β or worse, and F returns them as β or better. A and N again are the only examiners to use δ.

TABLE 112

337. PAPER IV (Political Theory)

Marks	Examiner									
	A	B	E	F	G	H	J	L	M	Q
	Number of Awards									
α	1	2	1	1	1	1	1			
Borderline	3		4	3	2				2	1
β	9	12	10	12	13	10	10	12	13	6
Borderline	2					3		4	1	5
γ		2	1			2	5			4
δ	1									
	16	16	16	16	16	16	16	16	16	16
Median	β?+ (12)	β?+ (12)	β?+ (12)	β / β?+ (11-12)	β?+ (12)	β / β?+ (11-12)	β+ (13)	$\beta-$ (9)	β+ (13)	$\beta\gamma$ (6)

L marks the scripts as β or worse, while F and G mark them as β or better. A is the only examiner to use δ.

338. We have the best basis for judging the differences between individual examiners if we consider the results of those who have marked three papers, i.e. A, B, F, H, J, K, L, and Q; Examiners A, B, F, H, J find clear α quality in some papers, whereas K, L, and Q never discover this quality.

Again, B, H, J, K, and Q discover clear γ quality in some papers, but A, F, and L do not, though A discovers δ quality in three papers. (A and N are the only examiners who award a δ.)

339. The averages (medians) of Q ($\gamma\beta$ for Paper I and $\beta\gamma$ for Paper III and Paper IV) differ fundamentally from the rest, all of which are in the range of β's. Of these examiners, B and L may be regarded as the extremes; their averages (medians) are set out below:—

	PAPER II	PAPER III	PAPER IV
B	$\beta+$?+ (14)	β+ / β?+ (13) (12)	β?+ (12)
L	β?− (10)	β / β?+ (11) (12)	$\beta-$ (9)

Q differs definitely from all the other examiners; and we get a fairer picture of the differences likely to occur in standard if we show the range of averages (medians) of the other examiners for the four papers set out below.

	PAPER			
	I	II	III	IV
Highest	(K) $\beta+$ (13)	(B) $\beta+$?+ (14)	(R) $\beta+$ (13)	(J & M) $\beta+$ (13)
Lowest	(P) $\beta-$ (9)	(J) $\beta-$ (9)	(N) β?$-$ (10)	(L) $\beta-$ (9)
Difference (Number of grades)	4	5	3	4

340. There is thus between these averages (medians) about four grades difference, from $\beta+$ to $\beta-$, corresponding to the familiar difference between II (i) and II (ii) of the Honours lists of some universities. We may say that there is between the standards of these examiners about half a class difference, even leaving Q out of account.

341. It is not surprising, if there are such differences between the averages (medians), that we should find much greater differences in the marking of individual scripts.

For Paper I, Table 105 shows that Candidate No. 13 was awarded α by Examiner O and $\gamma\beta$ by Examiner P, a range of 17 grades out of a possible range of 23. Q marks him $\beta\gamma$, but both D and K mark him $\alpha\beta$.

For Paper II, Table 106 shows that Candidate No. 8 gets $\alpha-$ from B and $\gamma+$ from J, a range of 16 grades, while Candidate No. 14 gets $\alpha-$ from B and $\gamma\beta$ from H, a range of 15 grades.

For Paper III, Table 107 shows that Candidate No. 9 gets α from A, and $\gamma+$ from B, a range of 18 grades ; while Candidate No. 3 gets α from R and $\beta\gamma$ from Q and N, a range of 16 grades.

For Paper IV, Table 108 shows that Candidate No. 8 gets $\alpha-$ from B and $\gamma+$ from J, a range of 16 grades.

342. These ranges are not affected by Q's low marking. Moreover, the average ranges (again leaving Q out of account) are as follows :—

For Paper I	8 grades
For Paper II	11 grades
For Paper III	10 grades
For Paper IV	9 grades

Thus on the average there is a whole class difference or thereabouts between the marks awarded by different examiners to the same script, since each class may be supposed to comprise about eight grades.

In no case does the same script get the same mark from all the examiners. The closest approach to equality is in judging the obviously very poor performance of Candidate No. 11 in Paper I ; he gets γ from two examiners and $\gamma-$ from the other three.

CHAPTER XI

A VIVA VOCE (INTERVIEW) EXAMINATION

343. *Object of the Investigation.*—The viva voce examination, not on a " subject " but of a general character, to test " alertness, intelligence, and general outlook," is an important element not only in Civil Service examinations but in interviews for the selection of candidates for public and private appointments generally.

It appeared, therefore, desirable to test the degree of consistency of two Boards of Examiners appointed to conduct an examination of this kind.

344. *Procedure.*—In order to secure a satisfactory basis for such an investigation, it was necessary to get together a suitable team of candidates.

The following conditions seemed desirable :—

(i) that the candidates should be approximately of the same age and have received the same kind of training ;

(ii) that the candidates should be provided with an adequate stimulus, not only to secure their presence but to make reasonably sure that they would treat the examination with the seriousness that is to be expected of candidates competing for an appointment ;

(iii) that the examiners should be provided with a suitable criterion by which the candidates were to be judged ;

(iv) that the examiners should be persons of experience, used to judging candidates by interview or viva voce examinations.

These conditions were fulfilled in the manner set out below.

345. The following notice was distributed to the universities and colleges of Great Britain and Northern Ireland :—

" The International Institute Examinations Enquiry offer a prize of £100 on the results of a viva voce examination. The examination will be

open to about fifteen candidates (men and women) selected from applicants studying at, or who have recently studied at, a University, and who are certified by the University or College authorities under whom they have been working to be suitable in their judgment as candidates for the Junior Grade of the Administrative Class (Home Civil Service). The examination for the prize will be held towards the beginning or end of the Easter Vacation, 1934, at a date to be announced subsequently. The age limits of the candidates, who must be British born subjects, are prescribed as follows :—

Candidates must have attained the age of twenty-one on the first day of August, 1933, and not have attained the age of twenty-three on the first day of August, 1933.

The examination will be in matters of general interest, not in matters of academic interest. It is intended to test the candidate's alertness, intelligence, and intellectual outlook. Each candidate must send in a record of his life and education. On the interview the examiners will judge of the value of the candidate's personality for the kind of career that the Home Civil Service offers.

Application must be made on forms to be obtained from the Director of the International Institute Examinations Enquiry, 1, Plowden Buildings, Temple, London, E.C.4, and must be returned through the authorities of the University or College concerned so as to reach the Director not later than 17th February, 1934.

Each candidate will be required to appear in London before two independent boards of examiners on the same day. Information will be supplied later as to the address at which the examination will be held. Travelling expenses up to an amount not exceeding 30s. will be paid to each candidate."

346. The number of candidates who applied was thirty. From these, sixteen candidates (twelve men and four women), with excellent university records, were selected for the purpose of the examination. The selection was made mainly, though not solely, on the ground of intellectual distinction, with a view to securing the kind of candidate who, if he or she competed for the Home Civil Service, might be expected to secure enough marks on the written work to be in the running for an appointment, and in whose case therefore the viva voce marks would be likely to act as a determining factor in the selection.

347. The candidates had received their university training in one or more of the following Universities and Colleges : Oxford, Cambridge, Glasgow, London, Bristol, University College, Nottingham, and University College, Southampton. Each candidate, on application, filled in the form marked A in the Appendix (p. 177), and, on selection, filled in the form marked B in the Appendix (p. 178).

348. Each examiner was supplied with copies of two items on Form A, (i) the confidential report from a tutor or other

authority, (ii) the candidate's own report on his life and education, and also with (iii) a complete copy of Form B as filled in by the candidate.

349. The following instructions were issued to the examiners :—

INTERNATIONAL INSTITUTE EXAMINATIONS ENQUIRY

Directions for the Meeting to be held on Tuesday, 27th March, at 9.45 a.m.

[The first three instructions give details as to time and place of meeting, interval for lunch, etc.]

(4) There will be two Boards of Examiners—Board I and Board II—each consisting of five examiners. The first business of each Board will be to elect their chairman, and to discuss any details of procedure other than those provided for in the scheme set out below.

(5) There will be sixteen candidates. These will be divided into two groups, Group A and Group B. Candidates in Group A will appear in alphabetical order first before Board I and then before Board II. Candidates in Group B will appear in alphabetical order first before Board II and then before Board I.

(6) Each candidate is to be examined *for not less than a quarter of an hour and not more than half an hour.*

(7) Particulars of each candidate, extracted from his[1] application, will be available for each examiner. The original application will be in the hands of the Chairman. The following is to be taken as the general direction with regard to the method of the viva voce examination.

The examination will be in matters of general interest, not in matters of academic interest ; it is intended to test the candidate's alertness, intelligence, and intellectual outlook. Each candidate has furnished a record of his life and education. On the interview and record the examiners will judge the value of the candidate's personality for the Home Civil Service.

The maximum mark is 300.

(8) In accordance with the letter of invitation to examiners, the following procedure will be adopted with regard to the recording of marks :

As soon as the viva voce examination of a candidate is over, *and before any discussion of his merits has taken place,* the Chairman will ask each of the examiners to write down his mark on the mark-sheet and he will also write down his own mark on his own mark-sheet. The Chairman will then ask the other examiners to state the marks so written down and will finally state his own mark so that each member of the Board may know what marks have been allotted in the first instance by the several members of the Board and be able to record them on his mark-sheet ; a discussion will then take place on the different marks proposed and the Chairman will record a mark representing the view of the Board as a whole, this mark being obtained either by agreement or, if

[1] The candidates and the Boards of Examiners will include women as well as men ; the masculine gender is used with reference to candidates and examiners only for the sake of simplicity.

that is impracticable, by taking an average of the marks allotted by the several examiners.

N.B.—The Chairman of each Board is requested to see that *the above arrangement is strictly observed*, as it is regarded as an essential feature of the Examination.

A suitable mark-sheet will be provided.

(9) Examiners are requested to sign and give in their mark-sheets to the Chairman of the Board.

350. The examination was held on 27 March, 1934, at the London School of Economics and Political Science, by kind permission of Sir William Beveridge, K.C.B., the Director, to whom the Committee desire to express their sincere obligation.

The examination began at about 10 a.m. and concluded in the late afternoon.

The names of the examiners, arranged in alphabetical order, were as follows :—

PROFESSOR ERNEST BARKER, Professor of Political Science, Cambridge, formerly Principal of King's College, London.

LADY VIOLET BONHAM-CARTER.

SIR FRANK DYSON, K.B.E., F.R.S., late Astronomer Royal.

MRS. MARY AGNES HAMILTON, formerly M.P. for Blackburn.

MISS H. REYNARD, M.A., Warden of King's College of Household and Social Science.

SIR HENRY RICHARDS, C.B., formerly Senior Chief Inspector, Board of Education.

PROFESSOR C. J. SISSON, Northcliffe Professor of Modern English Literature in the University of London.

MR. L. B. TURNER, Fellow of King's College and University Lecturer in Engineering, Cambridge.

DR. W. W. VAUGHAN, late Headmaster of Rugby.

Owing to the absence of one of the examiners (the Head of a College at one of the older Universities), who was unavoidably prevented from attending at the last moment, it was necessary to constitute one Board with only four examiners instead of five. But the statistical analysis which follows shows that there is no reason to think that the result was materially affected by this difference.

The examiners in Board I will be referred to as A, B, C, D, E ; those in Board II as F, G, H, I.[1]

351. At the end of the day each Board carefully reviewed its marks in order that the members might be sure that the marks

[1] These letters have no relation to the alphabetical order of the names of the examiners.

allotted translated correctly their impressions of the relative abilities of the candidates, since a prize was at stake.

352. The results, as set out in Table 113 below, are striking.

TABLE 113

Maximum Mark 300

No. of Candidate	*Board I* Initial marks awarded by the several examiners before discussion					Final Mark awarded by Board I	*Board II* Initial marks awarded by the several examiners before discussion				Final Mark awarded by Board II
	A	B	C	D	E		F	G	H	I	
1	130	120	150	150	100	120	190	210	210	240	212
2	260	260	250	260	250	260	200	210	200	140	190
3	130	140	150	150	120	130	190	180	185	160	175
4	240	220	170	210	280	230	250	280	250	260	255
5	230	210	170	230	190	210	260	210	210	250	232
6	230	150	190	190	180	180	220	260	260	220	250
7	210	180	150	225	200	200	270	280	280	230	270
8	250	260	170	250	200	240	230	200	225	240	224
9	230	230	180	230	230	230	270	220	165	250	220
10	210	250	180	230	180	210	230	250	260	200	235
11	170	210	170	250	200	210	250	225	220	250	236
12	220	240	170	220	250	230	250	270	200	210	232
13	120	120	150	120	100	120	160	180	180	190	177
14	230	230	170	180	230	210	230	280	220	260	247
15	240	220	170	200	200	220	200	210	190	180	195
16	180	100	160	180	240	170	220	200	150	190	175

353. The order in which the candidates were placed is shown in Table 114 below :—

TABLE 114

Candidate	*Board I* Mark	*Board II* Mark	*Board I* Order	*Board II* Order
1	120	212	15½	11
2	260	190	1	13
3	130	175	14	15½
4	230	255	4[1]	2
5	210	232	8½	7½
6	180	250	12	3
7	200	270	11	1
8	240	224	2	9
9	230	220	4[1]	10
10	210	235	8½	6
11	210	236	8½	5
12	230	232	4[1]	7½
13	120	177	15½	14
14	210	247	8½	4
15	220	193	6	12
16	170	175	13	15½

[1] The three candidates bracketed as equal after the first two candidates have been marked as "fourth" in order of merit in accordance with the usual practice in statistical tables.

354. *Award of the Prize.*—The prize was awarded to Candidate No. 4, Miss E. M. Francis, of Girton College, who was placed second by Board II and bracketed fourth by Board I.

The orders of merit of the two Boards are very different. The candidate placed first by Board I is placed thirteenth by Board II, and the candidate placed first by Board II is placed eleventh by Board I.

There were no cases of complete agreement in the marks assigned by the two Boards; the closest were the cases of Candidates Nos. 9, 12, 16, with 10, 2, 5 marks difference respectively. On the other hand there were extreme cases of disagreement; for Candidates Nos. 1, 2, 6 and 7, the differences were 92, 70, 70 and 70 marks respectively. The average difference is 37 marks. The extreme differences between the two Boards' estimates of the candidates' merits, amounting to 20 to 30 marks out of 100, and the average difference of about 12 marks out of 100, point to the unreliability of the interview test.

355. The coefficient of correlation between the marks of the two Boards is 0·41. This is comparatively small, and in view of the number of candidates involved cannot be considered "significant" in the usual sense. We must remember that the marks awarded are determined by two factors, the candidates and the Boards, and we must conclude that the different influences of the two Boards have been sufficient in this case almost to mask the common influence of the same set of candidates.

356. It is probable that the different questions asked of the candidates affect the marks finally awarded to the candidates. That the circumstances of the two interviews were different is apparent when we look at the individual assessments of the examiners.

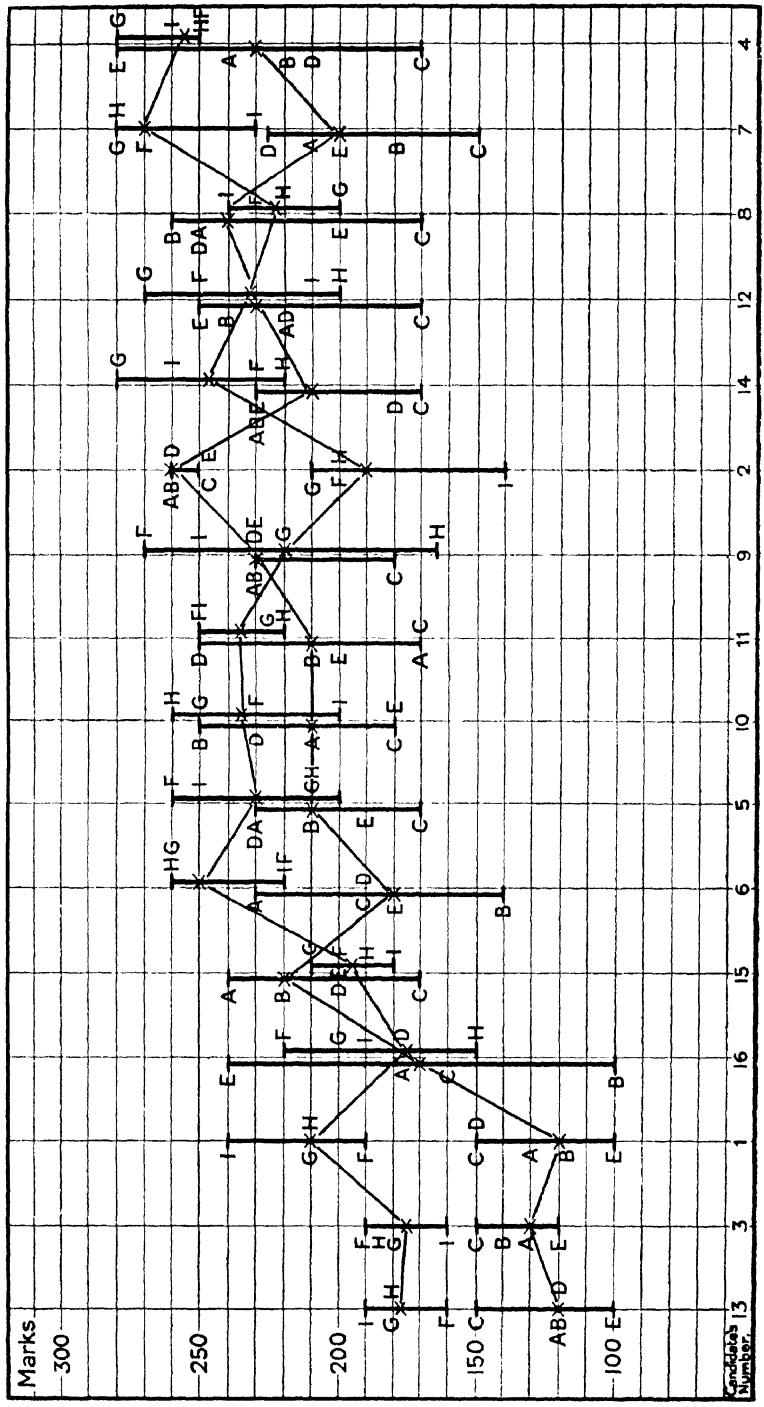
357. The chart on p. 174 shows in graphical form the full details of the results of the test.

For the candidates numbered 13, 3, 1, 2, and 7, the marks of the two Boards are entirely different—they do not overlap. The members of each Board were in agreement within different limits as to the merits of these candidates. Board I assessed the merits of Candidate No. 1 at 120, the individual examiners having awarded marks between 100 and 150; Board II assessed the candidate at 212, the individual examiners having awarded marks between 190 and 240.

358. Despite the identity of method used by the two Boards, as explained in the Postscript below, the actual evidence produced seems to have been so different that we might almost have supposed different candidates to have been examined.

VIVA VOCE EXAMINATION

The capital letters indicate the awards of the individual examiners. A thick vertical line connects the awards of the members of the same Board. A cross indicates the final award to each candidate by the Board.



In one respect there is a clear divergence between the results of the two Boards, since the average mark of Board I is 198, and the average mark of Board II is 220. The second Board on the whole gave higher assessments to the candidates.

359. Another striking case is that of Candidate No. 2. Board I gave him 260 marks, after very close agreement amongst the examiners as to his merits; Board II gave him 190 marks, the individual examiners' assessments ranging from 140 to 210.

360. The individual examiners' assessments show very close agreement in certain cases, the members of Board I agreeing within 10 marks in the case of Candidate No. 2, within 30 marks in the case of No. 3; and those of Board II within 30 marks in the case of Candidates Nos. 3, 4, 11, 13, 15.

But some of the marks are widely different. The different examiners of Board I gave Candidate No. 16, 180, 100, 160, 180, and 240 marks; they gave to Candidate No. 4, 240, 220, 170, 210, and 280 marks; the examiners of Board II gave to Candidate No. 9, 270, 220, 165, and 250 marks.

361. The average range of marks allotted by the various examiners to the several candidates was 51 in the case of Board II, and 69 in the case of Board I; but if we leave out of account the marks of Examiner C, which were consistently out of agreement with those of the rest of Board I, the average range for this Board is exactly the same as for Board II, namely, 51.

362. This agreement can be appreciated by means of the coefficient of correlation between the marks of the individual examiners and the final award of the whole Board. These are all significant when tested in the usual manner.

CORRELATION COEFFICIENTS

Board I

A	B	C	D	E
·91	·90	·63	·89	·84

Board II

F	G	H	I
·73	·86	·82	·72

363. The concurrent results show that, on the whole, each examiner on a Board was able to award a mark which was a fair reflection in most cases of the evidence placed before the Board, and therefore to agree with his colleagues as to the right mark. As one of us has pointed out in the postscript, the evidence placed before the two Boards was materially different, owing to the inherent character of an interview of this kind.

Postscript

364. I think that my impressions as an impartial observer of the proceedings of the two Boards (and one who has had experience in serving as an examiner at such viva voce examinations) may be of interest. The mode of approach of the two Boards seemed to me to be identical. They both appeared to me to succeed in securing the confidence of the candidates by tactful questioning and conversation carried on in nearly all cases as between equals. The candidates spoke with freedom and frankness.

As the two Boards met at the same time it was of course impossible for me to hear all the candidates examined by both. But I heard the two examinations of some of the candidates in regard to whom the differences of opinion were most striking. I came to the conclusion that, while the two Boards were equally skilful in cross-examining in such a way as to reveal the weaknesses of candidates, it was largely a matter of chance whether they struck on a topic in which a candidate felt so strongly that he was able to display his individuality. It would be impossible for me to quote the actual facts on which this opinion is based without revealing the personalities of the candidates concerned.

P. J. H.

APPENDIX TO CHAPTER XI

FORM A

International Institute Examinations Enquiry

Prize of £100

Application from intending Candidate

Name of Candidate (in block letters).....

Age in years and date of birth (accompanied by certificate of birth).....

Address (in block letters) to which letters should be forwarded, including vacation address, if any.....

.....

.....

.....

Name of University or College.....

Name of University or College Authority.....

Certificate of University or College

I, the undersigned, being.....(here state official position) certify that, in my own personal judgment, the candidate above-mentioned would be a suitable candidate for the Junior Grade of the Administrative Class (Home Civil Service); and I have appended hereto a confidential statement in regard to his (or her) character and intellectual attainments.

Date..... *Signature*.....

The candidate is to fill in on the opposite page a record of his (or her) life and education.

This letter, together with the confidential enclosure and certificate of birth, is to be forwarded, not by the candidate, but by the authorities of the College or University concerned, by registered post, to :—

The Director,
The International Institute Examinations Enquiry,
1, Plowden Buildings, Temple, London, E.C.4,

and must be delivered not later than 17th February, 1934.

Record of Applicant's Life and Education

(to be signed at the foot of the document)

FORM B

International Institute Examinations Enquiry
Viva Voce Examination

Form to be filled in by selected candidates

1. Name of Candidate	
2. Place of Birth, and state whether a natural born British subject	
3. Father's Name ,, Address ,, Profession or Trade (If deceased, give the last address, profession, etc.) Give place of father's birth and his nationality at birth Give place of mother's birth and her nationality at birth	
4. Name, in order, the Schools you have attended since the age of 12, giving addresses with dates of entering and leaving	
5. Give name or names of University or Universities, and dates of entering and leaving. State any degrees (with class obtained), honours or prizes you have obtained. Name your College or Colleges	
6. State any University or College colours, and any position of responsibility or distinction in University or College societies that you hold or have held	
7. If your time since leaving School is not fully accounted for by replies given above, account for the remainder here, with dates If you have had employers, state their names and addresses in full	
8. Signature and date	

PART II

By

E. C. Rhodes

CHAPTER XII

ON DIFFERENCES OF STANDARD AND RANDOM VARIATIONS

365. In Part I of this investigation, the marks allocated by a number of examiners to the work of a number of candidates at different kinds of written examinations have been presented and analysed up to a certain point. The object of Part II is to push the statistical analysis further.

In order to facilitate the discussion it is necessary to re-survey briefly the processes which lead up to the marks which are to be analysed.[1]

366. A final mark may be obtained by addition of many subsidiary marks ; for example, a paper may consist of a number of questions, each marked separately ; and again, each question may consist of a number of smaller parts, each marked separately.

Again, the process of marking may sometimes involve not only additions, but subtractions, a maximum for a question or part of a question being previously fixed, and marks taken off for specific mistakes. In an Essay paper, the attention of the examiners may be directed to a specific number of elements which are supposed to be discernible and each element may be separately marked, the total giving the final mark. On the other hand an Essay may be marked purely by impression, one mark only being awarded.

367. We shall find it convenient to use the phrase " a piece of work " as a general term, to describe either a set of scripts written by a candidate, containing the answers to a series of examination papers ; *or* a script written by a candidate, containing the answers to questions in one examination paper ;

[1] These processes are analysed in a somewhat different way by Professor Cyril Burt in a Memorandum following Part II (p. 245 below).

or a single answer to a question in the paper ; *or* a part of a single answer. Thus any mark used as the subject of our analysis may be regarded as awarded to " a piece of work," and this mark may be obtained by additions or subtractions of other marks, or may be an original mark. We shall refer to the smallest element to which marks are awarded, on any particular occasion, as the " unit piece of work."

368. An examiner, when assessing the value of a unit piece of work, may have a standard or model to which he refers. For instance, in Dictation an examiner would have the original passage of dictation, and in Arithmetic he would have the answers to simple sums before him.[1] In other cases the model piece of work may not be so easily available, but the examiner may have clearly defined instructions, how much to allot to a certain answer, how much to take off for a certain type of mistake, and so on. At other times there may be neither model nor precise instructions, but the examiner has in his own mind some sort of ideal answer.

369. Discrepancies may arise between two examiners' assessments of the same unit piece of work in a variety of ways. The candidate's writing may be poor or untidy ; in such a case, one examiner may read one meaning into what is written, another examiner may read something different. One examiner's idea of " perfection " may be different from that of the other. This kind of difficulty may be partly overcome by discussions beforehand, but in the essay type of question is never completely overcome. Even when examiners are expected to relate the piece of work to given models, their ideas of what are " like " or " unlike " the model may be different.

370. In addition, there is always the possibility of mistakes being made by inadvertence. An examiner may write 3, thinking of 5 marks ; or may write 7 and later read it as 1.

371. When a number of unit pieces of work submitted by many candidates are being examined by two examiners it may be that one examiner's powers of discrimination are not so well developed as the other's, or indeed the other examiner may imagine differences between the pieces of work which do not really exist.

372. When examiners are marking large numbers of scripts and the marking is spread over weeks, the possibility of fatigue

[1] Model essays have been used extensively in America by Professors Thorndike and Hillegas and others; they have been used in this country by Dr. W. Boyd in his *Measuring Devices in Composition, Spelling, and Arithmetic* (Harrap & Co.), 1924, and more recently by Dr. J. Perrie Williams, in *The Northamptonshire Composition Scale* (Harrap & Co.), 1933.

and boredom and state of health affecting the results must be considered. Where a model answer is in existence there is no possibility of the standard of achievement being altered, but there is the chance that the examiner's interpretation of words, phrases, symbols, etc., written in the script may change during the period of marking. Where there is no model to refer to, and where there is only a vague " ideal " in the mind of the examiner, it is obviously possible that the standard of reference may wobble.

373. When the marks awarded for unit pieces of work are summed, we may find that differences between examiners in respect of the units may cancel out because an examiner may on the whole be more generous in respect of one unit and more severe in respect of another. Such approximate cancellation may take place through inequalities in the marking of two different examiners. Thus a severe examiner may undermark one piece of a candidate's work and a generous examiner may overmark another piece. For instance, we may consider the following hypothetical case of two candidates examined by two examiners, A and B, who award the marks given below to two unit pieces of work.

	Qn. 1		Qn. 2		Total	
Max. Mark	20		20		40	
Examiner	A	B	A	B	A	B
Cand. 1	13	11	0	3	13	14
Cand. 2	11	8	3	5	14	13

Here Examiner A is more generous than B when marking Qn. 1, but more severe when marking Qn. 2. On the total marks, according to A, Candidate No. 2 is the better ; according to B, Candidate No. 1 is the better.

374. The examiners in the above illustration are consistent in their placing of these candidates in order, in regard to each of the two unit pieces of work, yet when the marks are totalled they place them in a different order.

375. The following illustration taken from the detailed marks of the English B paper in the Special Place examination is of interest (see paras. 194–207). The marks given are those awarded by Examiners H and J to the first ten candidates, for their work on Qns. 1, 2, 3, and 4 of this part of the English paper.

TABLE 115

	Questions											
	1		2		3		4		Total		Order	
Max. Mark	14		12		12		12		50			
Examiner	H	J	H	J	H	J	H	J	H	J	H	J
Candidate No. 1	8	9	9	8	11	8	10	10	38	35	2½	4
2	7	7	12	11	7	8	10	10	36	36	4	3
3	13	14	8	5	10	8	10	10	41	37	1	2
4	5	9	7	6	9	6	12	12	33	33	6	6½
5	10	11	6	6	8	8	0	0	24	25	10	9
6	2	4	6	5	7	7	12	12	27	28	8	8
7	3	3	8	6	9	8	6	6	26	23	9	10
8	6	8	9	5	8	8	12	12	35	33	5	6½
9	7	10	8	8	12	10	11	10	38	38	2½	1
10	8	12	4	8	8	4	10	10	30	34	7	5
Average	6·9	8·7	7·7	6·8	8·9	7·5	9·3	9·2	32·8	32·2		

376. Here the examiners are fairly consistent in their marking of the four unit pieces of work. Examiner H marks Qn. 1 on a lower scale than does Examiner J, but marks Qns. 2 and 3 on a higher scale. Only in the case of Candidate No. 10 is the mark of H lower than that of J for Qn. 2, and in the case of Candidate No. 2 for Qn. 3. Qn. 4 is marked by both examiners in the same way ; they only differ in regard to Candidate No. 9. The total marks differ by 4 in two cases (Candidates Nos. 3 and 10), by 3 marks in two cases (Candidates Nos. 1 and 7), by 2 marks in one case (Candidate No. 8), by 1 mark in two cases, and are exactly the same in three cases (Candidates Nos. 2, 4 and 9). The average marks are very nearly the same.

377. Examiner J gives six candidates 8 marks for Qn. 3, while the marks allotted by H for this piece of work are 11, 7, 10, 8, 9, 8 ; H certainly shows more sense of discrimination in regard to this question than does J.

378. Candidates Nos. 3 and 9 are interesting for comparison. Both examiners agree that Candidate No. 3 is better than Candidate No. 9 in regard to Qn. 1. In regard to Qn. 2 Examiner H places them equal, but Examiner J places Candidate No. 9 above Candidate No. 3. In regard to Qn. 3 both examiners agree in placing No. 3 below No. 9 ; and for Qn. 4 Examiner H gives Candidate No. 9 one more mark, but Examiner J gives them

the same mark. If the marks for Qns. 2, 3 and 4 are summed, we have the following results :—

	Examiner H	J
Candidate No. 3	28	23
,, ,, 9	31	28

Thus both examiners agree in placing Candidate No. 9 above Candidate No. 3. But when the marks for Qn. 1 are included, H places Candidate No. 3 above Candidate No. 9, and J reverses this order. The fact that Examiner H marked Qn. 1 rather more severely than did J meant that Candidate No. 9 got a rather low mark from H for his performance, and this has made the observed difference in the final result.

379. A further illustration taken from the same source shows how discrepancies are introduced even when the general levels of marking are the same. The marks in the following Table are those awarded by Examiners B and D to the same ten candidates as in the previous illustration.

TABLE 116

	Questions											
	1		2		3		4		Total		Order	
Max. Mark	14		12		12		12		50			
Examiner	B	D	B	D	B	D	B	D	B	D	B	D
Candidate No. 1	10	9	12	9	9	11	10	10	41	39	3½	5½
2	9	7	12	12	8	8	10	10	39	37	5	7
3	12	13	11	9	11	11	11	10	45	43	2	2½
4	8	9	11	10	10	10	12	12	41	41	3½	4
5	12	13	9	8	9	7	0	0	30	28	10	10
6	4	2	8	9	7	9	12	12	31	32	9	8
7	4	3	12	10	12	12	7	6	35	31	7	9
8	7	10	9	12	8	9	12	12	36	43	6	2½
9	11	11	12	12	12	11	11	10	46	44	1	1
10	8	12	4	9	11	8	10	10	33	39	8	5½
Average	8·5	8·9	10·0	10·0	9·7	9·6	9·5	9·2	37·7	37·7		

In this case the averages for each question and for the total are the same or very nearly the same. But the individual marks awarded by the two examiners are not always the same for each question, and the total marks differ by 7 in the case of Candidate No. 8, by 6 in the case of Candidate No. 10, by 4 in the case of

Candidate No. 7, by 2 in five cases (Candidates Nos. 1, 2, 3, 5 and 9), by 1 in the case of Candidate No. 6, and are the same in the case of Candidate No. 4. The orders of merit are different.

380. B gives a higher mark than D for Qn. 1 to four candidates, a lower mark to five candidates, and the same mark to one candidate. B gives a higher mark than D for Qn. 2 to five candidates, a lower mark to three candidates, and the same mark to two candidates. B gives a higher mark than D for Qn. 3 to three candidates, a lower mark to three candidates, and the same mark to four candidates. B and D give the same mark to seven candidates for Qn. 4, and B gives a higher mark than D to three candidates.

381. The marks awarded by the examiners to Candidates Nos. 7 and 10 are interesting. Both examiners agree in placing Candidate No. 7 below Candidate No. 10 in respect of Qns. 1 and 4, and in placing No. 10 below No. 7 in the case of Qns. 2 and 3. In the result B places 7 above 10, and D places 10 above 7. This is mainly because B gives Candidate No. 10 marks for Qns. 1 and 2 which are very much below those awarded by D. The superiority of Candidate No. 10 over Candidate No. 7 in respect of Qn. 1 is assessed by Examiner B at 4 marks, by Examiner D at 9 marks. The superiority of 7 over 10 in respect of Qn. 2 is assessed by B at 8 marks and by D at 1 mark. The superiority of 7 over 10 in respect of Qn. 3 is assessed by B at 1 mark, and by D at 4 marks, and the superiority of Candidate No. 10 over Candidate No. 7 is assessed by B at 3 marks and by D at 4 marks.

382. These differences between the judgments of the examiners in respect of the differences between the various pieces of work mean in the total a displacement of the order of these two candidates.

383. Let us look again at the marks awarded to answers to Qn. 2. Examiner D gives Candidates Nos. 1, 3, 6 and 10, 9 marks each ; Examiner B gives these candidates 12, 11, 8, 4 marks, 2 greater and 2 less than 9. Again, let us look at the marks awarded to answers to Qn. 3. Examiner D gives 11 marks each to Candidates Nos. 1, 3 and 9, while Examiner B gives these candidates 9, 11, and12 marks ; 1 less than, 1 equal to and 1 greater than 11. Further, while Examiner B gives 12 marks (the maximum) to Candidates Nos. 1, 2, 7 and 9 for answers to Qn. 2, D gives these candidates 9, 12, 10 and 12 marks.

384. Are the examiners introducing " errors of judgment " comparable to the errors of judgment which occur in all physical measurements ? If two observers are reading a thermometer

simultaneously, and estimating to tenths or twentieths of a degree, although they are using the same standard, their estimates may be different.

The fact that one examiner may give four candidates the same mark, and another give the same candidates 2 marks that are greater and 2 that are less, suggests that the processes are comparable; that in such a case the two examiners may be using the same standards, but that the comparison with the standards is made with different degrees of accuracy.

385. On the other hand, when the differences between the marks of two examiners are constantly of the same sign in dealing with different pieces of work, this suggests that the two examiners have all along in their minds different standards.

386. The total marks awarded by two examiners for a number of unit pieces of work may exhibit differences due to a combination of errors of judgment and differences of standard. We may for instance find that two examiners differ in the same sense in the marks allotted to *all* the unit pieces of work, and at the same time introduce errors of judgment of a random character. In such a case the constant difference between the examiners would cause a difference between the averages of the marks of a group of candidates. There would still remain the discrepancies in individual cases due to errors of judgment.

387. Also, we may have cases where one examiner, A, may have a higher standard for one piece of work than another examiner, B, but may have a lower standard than B for another piece of work. In such a case A would award on the average a lower mark than B for the first piece of work to all candidates, but would award a higher mark than B for the other piece of work to all candidates. Here again, in addition, errors of judgment may also enter into the marking. Variations of this particular kind in the standards of marking from one piece of work to another might tend to cancel out, and the average marks of A and B for a whole group of candidates might be the same, but discrepancies between the examiners' marks awarded to individual candidates would still persist, due to errors of judgment combined with those differences between the standards of marking of the unit piece of work which are not completely cancelled out in the case of the marks of individual candidates.

388. Thus we look for constant differences between examiners' marks indicated by differences between group averages, and for random variations indicated by the discrepancies which remain after allowance has been made for constant differences of standard.

The constant differences of standards of marking referred to

would only affect the average mark of a group of candidates, and would not affect their order of merit. The random variations would not affect the average mark of a group of candidates, but would alter the order of merit.

389. We shall consider an examiner who introduces into his marking random variations of a large order to be less precise in his marking than one whose marks contain less of this element, the perfect examiner being one who introduces no random variation into his marking.

390. Now let us suppose we had a " perfect " examiner, who could assign to each piece of work the " ideal " mark.[1] The ordinary examiner[2] would differ from him in two ways : his standard might not be the same, and he might introduce random variations into his marking.

391. We shall, by means of certain hypotheses, try to calculate approximately from the actual marks awarded to a number of scripts by a number of examiners, the appropriate ideal marks of the scripts.

392. We shall assume that the average verdict of a number of examiners is better than any of the single verdicts ; we might therefore use the simple average of the examiners' marks as the approximation to the ideal. But, in computing a simple average, each examiner's mark is allotted the same relative importance. A better approximation to the ideal mark will be obtained if a weighted average of the examiners' marks is used, the weights accorded to the several examiners being indicative of their relative precisions.

393. Suppose we consider as an illustration the marks awarded by the six examiners of Board I to the fifteen candidates in the Latin School Certificate investigation in our enquiry (see Table 12, para. 35), and find the simple average of each candidate's marks. These are shown in Table 117.

394. If we assume that these averages represent approximations to the ideal marks, then the differences between them and the original marks will indicate for each candidate how far the various examiners are departing from the ideal system. These differences are also shown in Table 117. It will be observed that these figures, in the case of Examiner A, are all negative, showing that A on the whole has more severe standards than the

[1] In what follows we use the term " ideal mark " to mean the mark that would be assigned by the " perfect examiner," as defined in para. 389 above, i.e. as one who introduces no random variations into his marking.

[2] Two ordinary examiners might be likened to two persons measuring the lengths of the same objects with tape measures made of materials of different elasticity and marked with scales having different zeros.

TABLE 117

(1)	Marks by examiners of Board I, Latin S.C. (2) (3) (4) (5) (6)					(7)	Approximation to ideal	Differences from approximation to ideal (8) (9) (10) (11) (12)					(13)	Constant Differences in Standards of Marking, and Random Variations (14) (15) (16) (17) (18)					(19)	Better approximation to ideal using weights
	Examiners						Average	Examiners						Examiners						
Cand.	A	B	C	D	E	F		A	B	C	D	E	F	A	B	C	D	E	F	
1	39	43	52	37	43	40	42	−3	+1	+10	−5	+1	−2	−4+1	+3−2	+7+3	−3−2	−2+3	0−2	41
2	39	44	50	43	43	46	44	−5	0	+ 6	−1	−1	+2	−4−1	+3−3	+7−1	−3+2	−2+1	0+2	44
3	44	51	55	47	46	46	48	−4	+3	+ 7	−1	−2	−2	−4+0	+3+0	+7+0	−3+2	−2+0	0−2	47
4	37	46	43	44	40	43	42	−5	+4	+ 1	+2	−2	+1	−4−1	+3+1	+7−6	−3+5	−2+0	0+1	42
5	38	47	55	35	43	45	44	−6	+3	+11	−9	−1	+1	−4−2	+3+0	+7+4	−3−6	−2+1	0+1	44
6	45	50	54	45	45	49	48	−3	+2	+ 6	−3	−3	+1	−4+1	+3−1	+7−1	−3+0	−2−1	0+1	48
7	42	52	51	45	44	46	47	−5	+5	+ 4	−2	−3	−1	−4−1	+3+2	+7−3	−3+1	−2−1	0−1	46
8	43	49	53	47	46	46	47	−4	+2	+ 6	0	−1	−1	−4+0	+3−1	+7−1	−3+3	−2+1	0−1	47
9	32	42	49	34	36	38	38	−6	+4	+11	−4	−2	0	−4−2	+3+1	+7+4	−3−1	−2+0	0+0	38
10	37	40	48	37	39	42	40	−3	0	+ 8	−3	−1	+2	−4+1	+3−3	+7+1	−3+0	−2+1	0+2	40
11	38	42	47	39	36	39	40	−2	+2	+ 7	−1	−4	−1	−4+2	+3−1	+7+0	−3+2	−2−2	0−1	40
12	40	44	50	41	36	42	42	−2	+2	+ 8	−1	−6	0	−4+2	+3−1	+7+1	−3+2	−2−4	0+0	42
13	38	43	50	36	34	41	40	−2	+3	+10	−4	−6	+1	−4+2	+3+0	+7+3	−3−1	−2−4	0+1	40
14	35	45	49	37	40	40	41	−6	+4	+ 8	−4	−1	−1	−4−2	+3+1	+7+1	−3−1	−2+1	0−1	40
15	32	38	41	28	34	34	34	−2	+4	+ 7	−6	0	0	−4+2	+3+1	+7+0	−3−3	−2+2	0+0	34
							Average	−4	+3	+ 7	−3	−2	0	Average of Squares of Random Variations (=Variance) 2·3	2·3	6·7	6·9	3·7	1·6	
														Weights 0·21	0·21	0·07	0·07	0·13	0·30	Total 1·00

ideal; that C's figures are all positive, showing that on the whole C is generous in his marking; and that F's figures are some positive and some negative, and some zero. The averages of the differences show to what extent the examiners consistently depart from the ideal.

395. For each examiner we can express the differences in the Table in terms of the consistent difference of standard plus or minus the "random variation." These differences are also shown. It will be observed (col. 14) that the random element of Examiner A is never greater than 2 marks; that of C (col. 16) is as much as 3, 4 and 6 marks; that of E (col. 18) is as much as 4 marks; and that of F (col. 19) is 2 marks in the case of only four candidates.

396. As we have already stated, we regard an examiner who introduces large random variations into his marking as less precise than one who introduces little variation of this nature into his marking. Consequently, on the evidence before us, we should regard A and F as more precise in their marking than C and D, and we should pay more regard to the marks awarded by the two former examiners than to those of the latter when we are estimating the ideal marks. We can do this by using a weighted average instead of a simple average as we have done.

397. We can obtain weights in the following manner. Let us find the variance of each examiner's random variations, this being the average of the squares of the random variations, and take as weights figures which are inversely proportional to these variances. In this way an examiner with large random variations and consequently with a large variance will have a small weight, and one with small random variations and a small variance will have a large weight. The weights thus calculated are shown in the Table, and we see that F is considered to be the most precise examiner and his figures are to be given the greatest weight in a more accurate calculation of the ideal set of marks, whereas C and D are to have the lowest weights. The method followed in the above is similar to that used by the physicist when he is estimating the relative importance of several measurements of the same quantity.

398. The revised ideal figures calculated by using these weights are shown in the last column of the Table, and it will be seen that in this particular case there is not a great deal of difference between the first approximation to the ideal marks, the simple averages, and the second approximation, the weighted averages.

399. Sir Philip Hartog in "Examinations and their Relation to Culture and Efficiency," p. 103, quotes Edgeworth (*Journal*

of the Royal Statistical Society, 1888, p. 599 *et seq.*, 1890, p. 644 *et seq.*), who " postulates ' that the true or standard mark of any piece of work is the average of the marks given by a large number of competent examiners equally proficient in the subject and instructed as to the character and purpose of the examination.' " Edgeworth uses " true or standard " where we use " ideal " and by taking the simple averages assumes that " competent examiners equally proficient, etc." will be equally precise examiners. This of course may not be the case. Sir Philip, then going on to describe Edgeworth's method, says " The general method adopted by Professor Edgeworth is to ascertain what this standard is in a number of typical cases, and to investigate the divergency from this standard or *error* of the marks allotted by individual examiners." That is to say, Edgeworth's method consists of finding the average of the marks allotted by a number of examiners to a piece of work, taking this as the true or standard mark, and then measuring the divergence of each examiner's mark from this. He would then regard these divergencies as a group of " errors " and, analysing them, would deduce the size of such " error " which might be anticipated to occur on a given occasion when any examiner (competent and proficient in his subject, etc.) assesses a piece of work by the award of a mark.

400. We part company with Edgeworth at this stage. We consider the problem of many examiners marking many scripts under conditions as far as possible similar to reality. We obtain sets of divergencies from " standard " (our " ideal "), one set for each examiner, and investigate each set separately, finding that with a number of examiners each set is differently constituted, and that the examiners are not using the same standards, nor are they equally precise as markers.

401. *The General Problem.*—We are given n pieces of work written by n candidates. These are marked by m examiners, A, B, C, ... , and the marks are as indicated in the scheme set out below :—

	Examiner				
Candidate	A	B	C	D	...
1	X_1	Y_1	Z_1	U_1	...
2	X_2	Y_2	Z_2	U_2	...
3	X_3	Y_3	Z_3	U_3	...
:	:	:	:	:	
t	X_t	Y_t	Z_t	U_t	...

Thus the tth script is awarded marks X_t, Y_t, Z_t by A, B, C respectively. We suppose that the ideal marks are Q_1, Q_2, Q_3, ... Q_t, ... and assume that

$$X_t = Q_t + A_t, Y_t = Q_t + B_t, Z_t = Q_t + C_t \ldots$$

$A_t, B_t, C_t, \ldots$ being the differences between the marks awarded by the examiners and the ideal mark, due to the examiners' personal peculiarities.

402. We know the X's, Y's, Z's ... for all values of t from 1 to n; we want to estimate the Q's for all these values of t, that is, we want to estimate the ideal marks, and we want at the same time to estimate the A's, B's, C's, ... so that we may estimate by how much the examiners diverge from the ideal in their assessment of candidates' merits as indicated by the scripts. A_t, B_t, C_t, ... we have assumed each to consist of two parts, the one a constant difference of standard from the ideal from script to script, the other a random variation of standard from script to script. We indicate this by writing $A_t = A + a_t$, $B_t = B + b_t$, $C_t = C + c_t$, ... where A, B, C, ... are the averages of the A's, B's, C's, ... for all the values of t. Thus

$$A = \frac{\text{sum } (A_t)}{n}, B = \frac{\text{sum } (B_t)}{n}, \ldots$$

403. The a's, b's, c's ... indicating the random variations of marking are sometimes +, sometimes −. We must have a measurement of the size of these random variations over the whole group of answers, and the measurement used is their standard deviation; thus, we will refer to the extent of random variation over the whole group by s_a, s_b, s_c, ... where

$$s_a = \sqrt{\frac{\text{sum } (a^2_t)}{n}}, s_b = \sqrt{\frac{\text{sum } (b^2_t)}{n}}, s_c = \sqrt{\frac{\text{sum } (c^2_t)}{n}}, \ldots$$

404. The A's as a group can then be characterised by referring to A and s_a, A indicating by how much the examiner deviates from the ideal standard consistently, and s_a indicating to what extent he introduces random variations in his marking from script to script. Similarly with the B's, C's, ... If s_a, s_b, s_c, ... are small, we infer that there is not much random element in the marking; the examiners are fairly consistent in their standards of marking from script to script. On the other hand, if s_a, s_b, s_c, are large, we infer that the examiners are not able to keep to their standards very well. If s_a is small and s_b is large, we should consider Examiner A a better examiner than B on

the grounds that he maintains his standard from script to script more consistently than does B; he is more precise as an examining machine.

405. The process of the analysis consists first in obtaining estimates of the values of s_a, s_b, s_c ..., in order to estimate the relative merit of the examiners' markings. Having estimated the relative merits of the examiners' results in this way, we attempt an estimate of the values of Q_t for t varying from 1 to n, i.e. we estimate the ideal marks for the different candidates. These estimates are obtained as weighted averages of X_t, Y_t, Z_t ..., where the weights assigned indicate numerically the relative merits of the examiners A, B, C, ..., these weights being derived from our knowledge of s_a, s_b, s_c ... .

406. Having obtained estimates of Q_t, we can by subtraction from X_t, Y_t, Z_t ... get estimates of A_t, B_t, C_t ..., and thus obtain $\bar{A}$, $\bar{B}$, $\bar{C}$, ... We thus split up X_t, Y_t, Z_t ... into their constituent parts Q_t, A_t, B_t, C_t ..., and split up A_t, B_t, C_t ... into their constituent parts $\bar{A}$,a_t; $\bar{B}$,b_t; $\bar{C}$,c_t ....

407. If we take averages for the whole n marks awarded for a piece of work, and denote the averages of the X_t's, Y_t's, Z_t's, for the n values of t by $\bar{X}$, $\bar{Y}$, $\bar{Z}$, ..., and similarly the average of the Q_t's by $\bar{Q}$, we may write $X_t = \bar{X} + x_t$, $Y_t = \bar{Y} + y_t$, $Z_t = \bar{Z} + z_t$ ..., $Q_t = \bar{Q} + q_t$, $A_t = \bar{A} + a_t$, $B_t = \bar{B} + b_t$, $C_t = \bar{C} + c_t$, where the small letters denote deviations from the averages, and we shall have

$$x_t = q_t + a_t, \quad y_t = q_t + b_t, \quad z_t = q_t + c_t \ldots .$$

408. Let us consider the pair $x_t = q_t + a_t$, $y_t = q_t + b_t$. The difference $x_t - y_t = a_t - b_t$ will give us $(x_t - y_t)^2 = (a_t - b_t)^2 = a_t^2 + b_t^2 - 2a_tb_t$. If we sum this for all values of t from 1 to n, we get

$$S(a_t^2) + S(b_t^2) - 2S(a_tb_t) = S(x_t - y_t)^2,$$

the right-hand side of this equation being known from the original data.

409. The left-hand side of this equation is concerned with a_t and b_t, which are the deviations of the A_t's and B_t's from their averages, that is a_t and b_t indicate for different values of t how the personal equations of the examiners are changing from time to time as one piece of work after another is being marked. Now if the markings of the two examiners are absolutely independent, there is no reason why the vagaries of marking of the one should be related to the vagaries of marking of the other. We should therefore expect that sometimes a positive a

will be associated with a positive b and sometimes with a negative b, and over the whole range of answers we should expect that $S(a_t b_t)$ would be small. We will definitely assume that this is zero, and that the equation above can be written

$$S(a_t^2) + S(b_t^2) = S(x_t - y_t)^2.$$

410. Similarly, if we consider the first and third examiners' marks we should have

$$S(a_t^2) + S(c_t^2) = S(x_t - z_t)^2.$$

Similarly, $$S(b_t^2) + S(c_t^2) = S(y_t - z_t)^2.$$

and so on.

We obtain, then, $\frac{m(m-1)}{2}$ equations of this kind, when we combine all possible pairs of the m examiners' markings. These $\frac{m(m-1)}{2}$ equations involve the m unknown quantities $S(a_t^2)$, $S(b_t^2)$, $S(c_t^2)$, ... We can from these obtain estimates for the latter by the method of least squares, which, in effect, means not assuming that such expressions as $S(a_t b_t)$ are exactly zero as stated above, para. 409, but that those values of $S(a_t^2)$, $S(b_t^2)$, ... , are taken as the best estimates of these unknowns, which make $[S(a_t b_t)]^2 + [S(a_t c_t)]^2 + \ldots$ as small as possible.

411. The equations from which $S(a_t^2)$, $S(b_t^2)$, ... are obtained are :—

$$(m-1)S(a_t^2) + S(b_t^2) + S(c_t^2) + \ldots = S(x_t - y_t)^2 + S(x_t - z_t)^2 + \ldots$$
$$S(a_t^2) + (m-1)S(b_t^2) + S(c_t^2) + \ldots = S(y_t - x_t)^2 + S(y_t - z_t)^2 + \ldots$$
.....

If we put $S(a_t^2) = a$, $S(b_t^2) = \beta$, $S(c_t^2) = \gamma$,
$S(x_t^2) = \xi$, $S(y_t^2) = \eta$, $S(z_t^2) = \zeta$,
$x_t + y_t + z_t + \ldots = p_t$, $S(p_t^2) = l$.

we can write these equations :—

$$(m-2)a + a + \beta + \gamma + \ldots = (m-2)\xi + \xi + \eta + \zeta + \ldots - 2Sx_t(y_t + z_t + \ldots)$$
$$= m\xi + \xi + \eta + \zeta + \ldots - 2S(x_t p_t),$$
$$(m-2)\beta + a + \beta + \gamma + \ldots = m\eta + \xi + \eta + \zeta + \ldots - 2S(y_t p_t).$$
.....

Adding gives

$$(2m-2)(a + \beta + \gamma + \ldots) = 2m(\xi + \eta + \zeta + \ldots) - 2S(p_t)^2,$$

or $$a + \beta + \gamma + \ldots = \frac{m}{m-1}(\xi + \eta + \zeta + \ldots) - \frac{l}{m-1}.$$

Thence

$$(m-2)\,\alpha = m\xi - \frac{1}{m-1}(\xi+\eta+\zeta+\ldots) + \frac{l}{m-1} - 2S(x_t p_t),$$

or
$$\alpha = \frac{m}{m-2}\xi - \frac{2}{m-2}S(x_t p_t) - \frac{1}{(m-1)(m-2)}(\xi+\eta+\zeta+\ldots) + \frac{1}{(m-1)(m-2)}l.$$

Similarly,

$$\beta = \frac{m}{m-2}\eta - \frac{2}{m-2}S(y_t p_t) - \frac{1}{(m-1)(m-2)}(\xi+\eta+\zeta+\ldots) + \frac{1}{(m-1)(m-2)}l$$

and so on.

Thus if there are seven examiners, $m = 7$, and we have

$$S(a_t^2) = \tfrac{7}{5}S(x_t^2) - \tfrac{2}{5}S(x_t p_t) - \tfrac{1}{30}(S(x_t^2) + S(y_t^2) + \ldots) + \tfrac{1}{30}S(p_t^2)$$

with similar expressions for $S(b_t^2)$, $S(c_t^2)$, ...

412. We can thus obtain from the original figures estimates of $S(a_t^2)$, $S(b_t^2)$, ... , and from these we can get the standard deviations of the individual examiners' personal equations. For instance, in the case of the first examiner, $A_1, A_2, \ldots A_t, \ldots A_n$ is a group of marks indicative of the divergences of this examiner from the ideal ; their average is A ; the deviations are $a_1, a_2, \ldots a_t, \ldots a_n$, and the standard deviation is the square root of $\frac{S(a_t^2)}{n}$. We denote these standard deviations by $s_a, s_b, s_c \ldots$.

413. We may therefore classify our examiners according to the size of these standard deviations, saying that the best examiner is that one with the least standard deviation and the worst examiner is that one with the greatest standard deviation.

414. We may go further. Since the examiners have each estimated the merit of an individual candidate's answer (say the tth), some, in a more precise fashion than others, as explained above, we may combine these estimates, X_t, Y_t, Z_t, ... to obtain Q_t, the ideal mark, just as when measurements of the same quantity are made by different instruments, an average can be taken as the best approximation to the unknown measurement of the quantity. The method of combination of X_t, Y_t, Z_t ... is to take a weighted average, the weights being proportional to the degrees of precision of the examiners, which are inversely

proportional to the squares of the standard deviations. Thus we take as the best available estimate of Q_t:—

$$Q_t = \frac{w_a X_t + w_b Y_t + w_c Z_t + \dots}{w_a + w_b + w_c + \dots}$$

where $w_a : w_b : w_c : \dots = \frac{1}{s_a^2} : \frac{1}{s_b^2} : \frac{1}{s_c^2} : \dots$

415. In this way we can obtain an estimate, for each answer, of the ideal mark. From the whole group of such estimates we can get the extent of the natural variation[1] amongst the candidates themselves, that is, we can get each q_t and thus obtain the standard deviation (s_q) of the group. This standard deviation may be used to indicate the extent of the natural variation in the group of candidates, which may be compared with the s_a, s_b, s_c, ... indicating the amount of variation introduced by the examiners.

[1] See footnote to para. 484.

NOTES TO CHAPTER XII

NOTE I

CONNECTION BETWEEN CORRELATION COEFFICIENTS AND THE SIZE OF THE RANDOM ELEMENT IN MARKING

416. Observations such as those which have been the subject of analysis in the foregoing pages are sometimes subjected to the correlation calculus, with the object of bringing out the degree of resemblance between marks awarded by different examiners to the same scripts. The coefficients of correlation thus worked out merely demonstrate in a simple way the discrepancies noted, and it is useful to show the connection between such coefficients and the extent to which the random variations enter into the marking.

417. We suppose that two examiners, A and B, mark a number of scripts. If their standards and methods of marking are identical the resulting marks will also be identical. But this in practice is not the case. A awards marks $X_1, X_2, \ldots X_n$ to the n scripts; B awards marks $Y_1, Y_2, \ldots Y_n$ to the scripts. We can suppose that there are ideal marks $Q_1, Q_2, \ldots Q_n$ to which $X_1, X_2 \ldots$; $Y_1, Y_2 \ldots$ are approximations. We may write $X_t = Q_t + A_t$; $Y_t = Q_t + B_t$ (for $t = 1$ to n); then the A's and B's indicate the examiners' divergences from the ideal. If we take averages for the whole group X, Y, Q, A, B, and deviations from these averages x_t, y_t, q_t, a_t, b_t, then a study of these deviations as a group will give us an indication of the peculiarities of the examiners on the evidence of the group marks as a whole. We have

$$x_t = q_t + a_t; \quad y_t = q_t + b_t.$$

418. We know the x's and y's, and, as described above, we can estimate the q's and a's and b's. s_q indicates the true variation in the whole group, s_a, s_b indicate the variations in the examiners' standards, i.e. the extent to which random marks enter into their estimates of the merits of a group of scripts.

419. The connection between the correlation between x and y and the ratios of s_a and s_b to s_q is indicated below.

$$r_{xy} = \frac{S(xy)}{\sqrt{S(x^2)S(y^2)}}$$

The numerator may be written $S(xy) = S(q^2) + S(qa) + S(qb) + S(ab)$, and if we assume that the variation in examiners' standards is absolutely random, $S(qa)$, $S(qb)$, $S(ab)$ should be exactly zero; so we can take $S(xy)$ as approximately equal to $S(q^2)$.

Similarly in the denominator $S(x^2) = S(q^2) + 2S(qa) + S(a^2)$, which is approximately $S(x^2) = S(q^2) + S(a^2)$.

Similarly $S(y^2)$ is approximately $S(q^2) + S(b^2)$.

420. Therefore r_{xy} is approximately

$$\frac{S(q^2)}{\sqrt{(S(q^2)+S(a^2))(S(q^2)+S(b^2))}} = \frac{1}{\sqrt{(1+S(a^2)/S(q^2))(1+S(b^2)/S(q^2))}}$$

Thus if $S(a^2)/S(q^2) = 0$ and $S(b^2)/S(q^2) = 0$, i.e. if the examiners show no personal vagaries when marking, $r_{xy} = 1$; there is an exact 1 to 1 correspondence in the marking. On the other hand, the larger $S(a^2)/S(q^2)$, $S(b^2)/S(q^2)$ are, the smaller is r_{xy}, i.e. the more the random element enters into the marking the less is the correlation between the results. s_a/s_q is, of course, the same as $\sqrt{S(a^2)/S(q^2)}$.

421. Now we can work out r_{xy} easily and affirm that there is a high degree of relationship between the marks, the random element being of small extent, if, for instance, $r = \cdot 9$; or if $r = \cdot 5$ we can say that the influence of the random element is more pronounced; but we can get no further than this purely adjectival description.

422. We have attempted to go further. We have tried to measure (approximately at any rate) the actual extent of the random element: consequently our work includes what would be done by a mere calculation of correlation coefficients and more besides.

423. We can construct a Table such as that below, showing values of r_{xy} for different combinations of s_a/s_q, s_b/s_q.

Values of s_a/s_q	0	·2	·4	·6	·8	1·0	1·2
Values of s_b/s_q	Values of r						
0	1	·980	·929	·858	·781	·707	·640
·2	·980	·962	·910	·841	·765	·695	·628
·4	·929	·910	·862	·796	·725	·656	·594
·6	·858	·841	·796	·736	·670	·607	·549
·8	·781	·765	·725	·670	·610	·552	·500
1·0	·707	·695	·656	·607	·552	·500	·452
1·2	·640	·628	·594	·549	·500	·452	·409

Thus we could infer, given $r = \cdot 5$ (say), and assuming that s_a/s_q, s_b/s_q were about the same, that each ratio was approximately unity, i.e. that the extent of the random element in each examiner's marks was about the same as the extent of the natural variation in the group of candidates. Alternatively, knowing s_a/s_q, s_b/s_q, we can infer the size of r, e.g. if these are about $\cdot 2$, r is $\cdot 96$.

NOTE II

ON THE ASSUMPTION THAT THERE IS NO CONNECTION BETWEEN RANDOM VARIATIONS OF DIFFERENT EXAMINERS

424. There is one assumption in Chapter XII to which further attention is necessary. We have assumed that if two examiners are marking a piece of work the random variations of the two examiners will not bear any relation to one another. Now *prima facie* we might expect that if the ideal mark of a piece of work were 100 per cent., and two examiners' standards of marking were perfect, any random variation introduced by an examiner would be negative, and similarly at the other end of the scale, with a piece of work worth 0 marks in the ideal scale, any random variation would be positive.

425. We have avoided the difficulty in some of our investigations by confining our attentions to average pieces of work. We have also found that, at any rate in the upper reaches of the scale of marks, experience shows that our *a priori* reasoning is wrong. In the Mathematical Honours investigation we found that in the case of one candidate (No. 18) the maximum mark was exceeded by two examiners, D and F, who awarded 308 and 303 (maximum 300) (see Table, para. 518). In most of our investigations our main concern has been the effect of the discrepancies of marks on awards such as Credit, Pass, Failure, or First, Second, Third Classes, and naturally our attention has been mainly directed to the borderline cases. In all these our assumption is reasonably justified.

426. More refined assumptions which would allow for difficulties of this nature might be introduced in order to arrive at better approximations to the ideal marks, but it was not considered that much would be gained, at this stage, by proceeding further in this direction.

NOTE III

ON THE SPREADING OF IDEAL MARKS

427. In Chapter XII it has been assumed that all examiners attempt to find the same ideal mark. There is the possibility that this may not be the case, that some examiners may " spread " the ideal marks more than others. Thus, instead of assuming as has been done above (para. 401) that

$$X_t = Q_t + A_t,$$

we may assume $$X_t = r_a Q_t + A_t,$$

where r_a is a multiplier peculiar to Examiner A. This change would admit the possibility, in the analysis, of each examiner having a different set of ideal marks. This refinement is dealt with in greater detail in my Memorandum on " A Second Approximation for the Determination of Ideal Marks and Random Variations " (pp. 315–324 below), where comparisons are also made between results obtained by using the two methods of analysis on the same sets of data.

CHAPTER XIII

RESULTS OF APPLYING THE METHOD OF ANALYSIS TO THE DATA OF THE INVESTIGATIONS

Section 1.—*School Certificate History* (see paras. 1–27)

428. The marks awarded by Examiners J and Q in the first investigation may fruitfully be compared with the object of showing the occurrence of random variations; and those of B and H, with the object of showing differences due to differences of standards of marking.

428a. The average mark of Examiners J and Q is the same, 46·8, but the marks awarded by these two examiners to individual candidates are, as shown below, very different.

Candidate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Examiner															
J	31	31	43	52	47	41	58	36	43	57	55	70	38	47	53
Q	50	42	46	43	48	47	51	45	37	40	52	62	45	41	53
Difference	−19	−11	−3	+9	−1	−6	+7	−9	+6	+17	+3	+8	−7	+6	0

(Extract from Table 2, para. 7.)

In the case of seven candidates, J's marks are less than Q's; in the case of another seven, J's marks are greater than Q's; and in one case the marks are the same. In the case of Candidate No. 1 there is a difference of 19 marks, which actually means the difference between Failure and Credit.

429. A difference of standard of marking will lead to one examiner's marks awarded to a set of scripts being on the whole lower or higher than those awarded by another examiner, and will thus tend to change the numbers of Credits, Passes and Failures as between the two examiners. The effect of such differences of standard can be observed easily by a comparison of average marks. For example, in the first investigation, Examiner H's average was 51·0, and that of B was 34·6, a difference of 16·4 marks (out of 96). H marked on the whole much

more generously than B. The marks for each candidate are as follows (see Table 2, para. 7) :—

Candidate	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Examiner															
H	48	46	49	52	47	51	58	41	46	55	49	61	44	53	65
B	33	27	31	40	34	37	48	22	27	41	29	40	25	41	44
Difference	15	19	18	12	13	14	10	19	19	14	20	21	19	12	21

H awarded 7 Credits and 8 Pass marks; B awarded 6 Pass marks and 9 marks of Failure.

430. We may illustrate the results of our procedure by quoting the appropriate components into which the marks are split up in the cases of Examiners B and H in the first investigation.

TABLE 118

		EXAMINER B			EXAMINER H		
Candidate	Ideal	Constant Difference	Random Variations	Original Marks	Constant Difference	Random Variations	Original Marks
1	42	−10·1	+1·1	33	+6·3	−0·3	48
2	38	−10·1	−0·9	27	+6·3	+1·7	46
3	44	−10·1	−2·9	31	+6·3	−1·3	49
4	48	−10·1	+2·1	40	+6·3	−2·3	52
5	43	−10·1	+1·1	34	+6·3	−2·3	47
6	47	−10·1	+0·1	37	+6·3	−2·3	51
7	52	−10·1	+6·1	48	+6·3	−0·3	58
8	38	−10·1	−5·9	22	+6·3	−3·3	41
9	36	−10·1	+1·1	27	+6·3	+3·7	46
10	46	−10·1	+5·1	41	+6·3	+2·7	55
11	44	−10·1	−4·9	29	+6·3	−1·3	49
12	55	−10·1	−4·9	40	+6·3	−0·3	61
13	39	−10·1	−3·9	25	+6·3	−1·3	44
14	43	−10·1	+8·1	41	+6·3	+3·7	53
15	56	−10·1	−1·9	44	+6·3	+2·7	65
	Averages		Standard Deviation		Averages		Standard Deviation	Average
	44·7	−10·1	4·1		34·6	+6·3	2·3	51·0

The random element is larger in the case of Examiner B than with Examiner H, and this is reflected in the higher standard deviation of his random variations.

431. The presence of this random element naturally affects the order of merit of the candidates, whereas the effect of a constant difference between standards of marking obviously has no such influence. The following shows the order of merit of the candidates, according to the ideal marks, B's marks, and H's marks.

TABLE 119

Candidate	Ideal	Examiner	
		B	H
1	11	9	10
2	13½	12½	12½
3	7½	10	8½
4	4	5½	6
5	9½	8	11
6	5	7	7
7	3	1	3
8	13½	15	15
9	15	12½	12½
10	6	3½	4
11	7½	11	8½
12	2	5½	2
13	12	14	14
14	9½	3½	5
15	1	2	1

432. The ideal marks obtained from the examiners' marks at the two investigations are given below :—

TABLE 120

Column	Ideal Marks (1)	(2)	(3)	Order of Merit (4)	(5)	(6)
Candidate	1st Investigation	2nd Investigation	Difference (1) - (2)	1st Investigation	2nd Investigation	Difference (4) - (5)
1	42 (P)	45 (P)	−3	11	5½	5½
2	38 (F)	36 (F)	+2	13½	14	½
3	44 (P)	42 (P)	+2	7½	9	1½
4	48 (P)	45 (P)	+3	4	5½	1½
5	43 (P)	43 (P)	0	9½	7½	2
6	47 (P)	49 (P)	−2	5	4	1
7	52 (C)	51 (C)	+1	3	1	2
8	38 (F)	37 (F)	+1	13½	12½	1
9	36 (F)	32 (F)	+4	15	15	0
10	46 (P)	43 (P)	+3	6	7½	1½
11	44 (P)	41 (P)	+3	7½	10	2½
12	55 (C)	50 (C)	+5	2	2	0
13	39 (F)	37 (F)	+2	12	12½	½
14	43 (P)	40 (P)	+3	9½	11	1½
15	56 (C)	49 (P)	+7	1	3	2
Average	44·7	42·7	+2·0			
Standard Deviations	5·9	5·5				

There is some difference between the two sets of ideal marks due to the fact that on the whole the examiners marked more severely on the second occasion than the first. There is as much

as 7 marks difference in the ideal mark of Candidate No. 15. If the ideal marks had formed the basis of classification, this candidate is the only one whose award would be different on the two occasions ; he would only " pass " at the second investigation, whereas he receives a " credit " at the first.

433. The standard deviations of the two sets of ideal marks are nearly the same, and are not very large. This is due to the fact that the group of candidates was a mediocre one.

434. The order of the candidates also changes with the change from the first set of ideal marks to the second, the greatest change being a movement in the case of Candidate No. 1 from eleventh place to equal fifth.

435. The constant differences of standard of marking, being the differences between the averages of the examiners' original marks and the averages of the ideal marks, are given below for both investigations.

CONSTANT DIFFERENCES OF STANDARD OF MARKING FROM IDEAL

Examiner	1st Investigation	2nd Investigation
A	−3·5	—
B	−10·1	−4·7
C	+0·4	+4·4
D	−2·5	−8·6
E	−1·0	+2·0
F	+2·1	−3·6
G	+3·3	+8·6
H	+6·3	+6·0
J	+2·1	+1·5
K	+5·3	+2·3
L	−6·0	−5·8
M	+0·5	+1·7
N	−5·9	+0·4
P	−2·1	+0·6
Q	+2·1	−1·2

Examiners B, D, L mark severely on both occasions, Examiners C, G, H, J, K, M mark generously on both occasions. The other examiners (except A)—E, F, N, P, Q—changed the signs of the differences between their average marks and the ideal averages. Some examiners, E, P, Q, did not differ greatly from the ideal.

In some cases, where there was considerable difference of standard from the ideal, this was maintained, e.g. with Examiners B, H, and L.

436. The standard deviations indicating the extent of the random element in marking at the two investigations are given below.

TABLE 121

Examiner	Standard Deviation of Random Variations		Order of Examiners according to size of Standard Deviation (Smallest S. D. placed first)[1]	
	1st Investigation	2nd Investigation	1st Investigation	2nd Investigation
A	4·9	—	—	—
B	4·1	5·5	4	8
C	7·0	7·0	10	12
D	4·4	5·2	5	6
E	7·2	8·0	13	13½
F	7·3	5·0	14	5
G	7·1	8·0	11½	13½
H	2·3	4·2	1	3
J	7·1	6·3	11½	11
K	4·6	5·4	6½	7
L	3·6	4·9	3	4
M	3·2	3·1	2	1
N	5·1	3·9	8	2
P	5·8	5·6	9	9
Q	4·6	6·1	6½	10
Standard Deviations of Ideal Marks	5·9	5·5		

The two sets of figures in the above Table of Standard Deviations are roughly of the same size, and the columns showing the order of the examiners according to this criterion show considerable resemblance. At the second investigation, those examiners with the smaller random variations at the first marking allot marks which again have the smaller random variations on the whole, the correlation between the two orders above being 0·66. As far as can be judged from this investigation, the examiners show some consistency in the extent of their random variations on two different occasions.

437. The standard deviation may be considered to indicate that if a candidate's ideal mark is, say, 50, an examiner with a random variation indicated by a standard deviation of 2·3 (say) would award a mark probably within a range of 4½ (twice the standard deviation) on either side of 50, i.e. his mark would probably be somewhere between 45½ and 54½. Thus on one occasion he may give 51 marks to such a script, on another 48 marks, on another 53 marks. Some of these standard deviations are quite high (over 7 marks), indicating that an examiner with such a loose standard of marking may award, instead of 50 marks, a mark somewhere in the range 35 to 65. Now in this kind of

[1] Examiner A is omitted from this half of the Table as he did not take part in the second investigation.

examination this range of marks would include the borderline marks for Pass and Credit. Thus a candidate who is possibly worthy of a Credit may actually achieve a Pass only, or even be dubbed a Failure, or may succeed in being given a mark of Credit instead of a Pass.

438. The extent of the variability amongst the candidates, due to their difference in ability to answer questions in this subject, as judged from the ideal marks, was 5·9 in the first investigation, and 5·5 in the second. The standard deviations of the random variations are in the case of many examiners of this order of size, and it is quite conceivable that the different standards of marking of the examiners combined with the random variations which are likely to occur, in view of the sizes of the standard deviations, would result in all these candidates being awarded exactly the same mark on some occasion. Actually this is what happened when the scripts were first marked for the Examining Authority. As stated in para. 2 (i), the scripts all received the same " middling " mark.

Section 2.—School Certificate French (see paras. 58–93)

439. The two sets of ideal marks for the two Boards are given below, together with the differences between them, and the respective places in the orders of merit, with their differences. We also show in this Table the actual orders of merit for the two Boards, with division lines indicating those candidates who would qualify for Distinction, Credit, Pass, Failure, if the ideal marks formed the basis of such awards.

440. The averages of the ideal marks are practically the same (47·6 and 48·0). The ideal marks of individual candidates differ by as much as 8 (Candidate No. 40), and in comparatively few cases are the marks the same. Thus there is no doubt that two boards of examiners will produce different ideal marks.[1] The orders of merit are also different, but the first three candidates and the last two are the same in both sets.

441. The last two columns in the Table show the candidates in order of merit, the division lines indicating the class groups, Distinction, Credit, Pass, Failure. If the ideal marks had determined the awards of classes, it is evident that the two sets of ideal marks would have led to substantially the same results, but for the fact that the two Boards had different limiting marks for the various classes. Thus we note that the first six in order in the list of ideal marks of Board II include five of those in

[1] It is course to be remembered that the calculated ideal marks shown in Table 122 and similar tables in Part II, are only approximations to the ideal marks, as defined theoretically.

TABLE 122

Ideal Marks

Cand. No.	Ideal Marks: Board I	Ideal Marks: Board II	Ideal Marks: Difference (I − II)	Order: Board I	Order: Board II	Order: Difference (I − II)	Candidates Arranged in Order of Merit: Board I Candidate No.	Candidates Arranged in Order of Merit: Board II Candidate No.	
1	63	65	−2	9	8½	+0½	48	48	
2	51	51	0	22½	26½	−4	36	36	D
3	48	50	−2	30½	29	+1½	D 46	46	
4	42	40	+2	34	37	−3	19	19	
5	22	17	+5	47	47	0	41	25	
6	53	58	−5	17½	15	+2½	25	18	
7	61	65	−4	11	8½	+2½	18	41	
8	34	40	−6	41½	37	+4½	12	1	
9	14	8	+6	49	49	0	1	7	
10	57	59	−2	12	13	−1	21	21	
11	49	51	−2	28	26½	+1½	7	12	
12	64	61	+3	8	11	−3	10	10	
13	50	53	−3	25	22	+3	44	44	
14	33	28	+5	43	43½	−0½	50	45	
15	44	48	−4	32	32	0	24	6	C
16	35	32	+3	40	42	−2	47	47	
17	6	3	+3	50	50	0	C 6	50	
18	65	67	−2	7	6	+1	22	31	
19	69	69	0	4	4½	−0½	23	23	
20	39	41	−2	37	35	+2	43	24	
21	62	62	0	10	10	0	45	43	
22	53	51	+2	17½	26½	−9	2	13	
23	52	54	−2	20	20	0	33	33	
24	54	54	0	15½	20	−4½	13	30	
25	67	69	−2	6	4½	+1½	34	11	
26	38	38	0	39	39	0	42	22	
27	34	35	−1	41½	41	+0½	11	2	
28	39	43	−4	37	33½	+3½	30	42	
29	39	40	−1	37	37	0	31	3	
30	49	52	−3	28	23½	+4½	3	34	
31	49	55	−6	28	18	+10	35	35	
32	42	43	−1	34	33½	+0½	15	15	
33	51	52	−1	22½	23½	−1	4	32	
34	50	49	+1	25	30½	−5½	32	28	
35	48	49	−1	30½	30½	0	39	20	P
36	73	72	+1	2	2	0	P 20	8	
37	19	21	−2	48	46	+2	28	4	
38	31	28	+3	44	43½	+0½	29	29	
39	42	36	+6	34	40	−6	26	26	
40	23	15	+8	45½	48	−2½	16	39	
41	68	66	+2	5	7	−2	8	27	
42	50	51	−1	25	26½	−1½	27	16	
43	52	54	−2	20	20	0	14	14	
44	56	59	−3	13½	13	+0½	38	38	
45	52	59	−7	20	13	+7	40	49	
46	72	70	+2	3	3	0	49	37	
47	54	57	−3	15½	16	−0½	F 5	5	F
48	82	80	+2	1	1	0	37	40	
49	23	22	+1	45½	45	+0½	9	9	
50	56	56	0	13½	17	−3½	17	17	
Average	47·6	48·0	−0·4						

the first six in the list of ideal marks of Board I. Candidate No. 41, who is fifth on the list of Board I is seventh on the list of Board II; and Candidate No. 18 who is sixth on the list of Board I is seventh on the list of Board II.

We note that if the borderlines defining Credit of Board II were moved down three places, the same candidates would be included in the Credit class by both Boards, with the exception of Candidates Nos. 18 and No. 41, noted above. If the Pass–Failure line were moved down two places for Board II, the only complication would be the placing of Candidate 38. It is apparent, then, that one of the main differences between the two boards as to the awards of the various classes is due to the fact that they have used different marks to indicate the borderlines between these classes.

442. The ideal distribution of awards would be as follows, adopting for each Board its own limits for the different categories.

	Board I	*Board II*
Failures	7	9
Passes	11	12
Credits	26	26
Distinctions	6	3
Total	50	50

443. The standard deviations of the random element in the individual examiners' final marks for the whole subject are shown below, together with the standard deviations of the two sets of ideal marks :—

Board I		*Board II*	
Examiner	Standard Deviation of Random Variations	Examiner	Standard Deviation of Random Variations
A	3·8	G	1·8
B	2·5	H	3·7
C	2·5	J	2·7
D	2·7	K	3·1
E	2·4	L	2·1
F	3·3	M	2·5
Standard Deviation of ideal marks	15·5	Standard Deviation of ideal marks	16·9
Maximum	100	Maximum	100

The extent of the random element is small compared with the amount of natural variation amongst the candidates, in the case of both sets of examiners. The detailed instructions to examiners naturally lead to precision of marking.

TABLE 123

Cand.	Board I Marks Ideal	Board I Marks Exr. C	Board I Difference between C and Ideal	Board I Awards Ideal	Board I Awards Exr. C	Board II Marks Ideal	Board II Marks Exr. J	Board II Difference between J and Ideal	Board II Awards Ideal	Board II Awards Exr. J
1	63	65	+2	C	C	65	68	+3	C	C
2	51	51	0	C	C	51	53	+2	C	C
3	48	50	+2	C	C	50	54	+4	C	C
4	42	46	+4	P	C	40	43	+3	P	P
5	22	25	+3	F	F	17	20	+3	F	F
6	53	53	0	C	C	58	58	0	C	C
7	61	62	+1	C	C	65	64	−1	C	C
8	34	37	+3	P	P	40	36	−4	P	P
9	14	15	+1	F	F	8	10	+2	F	F
10	57	59	+2	C	C	59	58	−1	C	C
11	49	50	+1	C	C	51	56	+5	C	C
12	64	61	−3	C	C	61	62	+1	C	C
13	50	51	+1	C	C	53	50	−3	C	C
14	33	32	−1	P	F	28	25	−3	F	F
15	44	42	−2	C	P	48	50	+2	P	C
16	35	32	−3	P	F	32	31	−1	F	F
17	6	6	0	F	F	3	0	−3	F	F
18	65	62	−3	C	C	67	70	+3	C	D
19	69	68	−1	D	D	69	73	+4	C	D
20	39	35	−4	P	P	41	42	+1	P	P
21	62	68	+6	C	D	62	59	−3	C	C
22	53	53	0	C	C	51	56	+5	C	C
23	52	50	−2	C	C	54	53	−1	C	C
24	54	54	0	C	C	54	50	−4	C	C
25	67	68	+1	D	D	69	70	+1	C	D
26	38	38	0	P	P	38	39	+1	P	P
27	34	30	−4	P	F	35	37	+2	P	P
28	39	40	+1	P	P	43	43	0	P	P
29	39	39	0	P	P	40	36	−4	P	P
30	49	49	0	C	C	52	47	−5	C	P
31	49	44	−5	C	C	55	57	+2	C	C
32	42	42	0	P	P	43	43	0	P	P
33	51	47	−4	C	C	52	50	−2	C	C
34	50	49	−1	C	C	49	49	0	P	P
35	48	51	+3	C	C	49	51	+2	P	C
36	73	72	−1	D	D	72	71	−1	D	D
37	19	18	−1	F	F	21	21	0	F	F
38	31	28	−3	F	F	28	26	−2	F	F
39	42	45	+3	P	C	36	30	−6	P	F
40	23	24	+1	F	F	15	14	−1	F	F
41	68	65	−3	D	C	66	65	−1	C	C
42	50	48	−2	C	C	51	50	−1	C	C
43	52	45	−7	C	C	54	52	−2	C	C
44	56	56	0	C	C	59	60	+1	C	C
45	52	55	+3	C	C	59	65	+6	C	C
46	72	70	−2	D	D	70	73	+3	D	D
47	54	53	−1	C	C	57	54	−3	C	C
48	82	83	+1	D	D	80	80	0	D	D
49	23	26	+3	F	F	22	22	0	F	F
50	56	54	−2	C	C	56	56	0	C	C

444. It is interesting to note the effect of the random element, by comparing Examiner C's marks with the ideal marks of Board I, the difference between C's average and the ideal average being negligible, and by comparing Examiner J's marks with the ideal marks of Board II, the difference between J's marks and the ideal average of Board II again being negligible.

These two sets of marks are given on page 207, together with the classified results.

445. The candidates whose class is affected by the random element in Examiner C's marking are Nos. 4, 14, 15, 16, 21, 27, 39 and 41, eight in all. The details are as follows :—

Candidate	Difference between C's mark and ideal	Ideal Award	C's Award	Raised + Lowered −
4	+4	P	C	+
14	−1	P	F	−
15	−2	C	P	−
16	−3	P	F	−
21	+6	C	D	+
27	−4	P	F	−
39	+3	P	C	+
41	−3	D	C	−

Thus a small difference of 1, 2 or 3 marks has the effect of making a difference of award in five cases.

Similarly the candidates whose award is affected by the random element in Examiner J's marking are Nos. 15, 18, 19, 25, 30, 35 and 39, seven in all. The details are as follows :—

Candidate	Difference between J's mark and ideal	Ideal Award	J's Award	Raised + Lowered −
15	+2	P	C	+
18	+3	C	D	+
19	+4	C	D	+
25	+1	C	D	+
30	−5	C	P	−
35	+2	P	C	+
39	−6	P	F	−

Again a small difference of 1, 2 or 3 marks has the effect of making a difference of award in four cases.

These two illustrations are typical of the effect of the random element on the results. In each case the random element is fairly small (a standard deviation of about $2\frac{1}{2}$ marks out of 100). In one case eight candidates, and in the other case seven candidates out of fifty have their award altered owing to the presence of the random element in the examiner's marks.

446. The differences between the standards of the examiners in marking the various questions have already received attention.

The average range of marks for the questions has also been dealt with (see paras. 82–93). The range is affected by the differences of standard of marking, but it is also affected by the random element in marking. It is interesting to consider the size of the random element introduced on the average in the marking of individual questions. The following Table shows the average of the examiners' standard deviations of random variations for the various questions.

TABLE 124

Question	Paper I					Paper II	
	Dictation	2a	2b	2c	Style	1	2
Board I	1·06	1·11	1·32	1·25	—	1·48	1·84
% *of Max.*	*9·7*	*7·4*	*6·6*	*8·2*	—	*4·9*	*9·2*
Board II	1·01	2·23	2·18	1·38	1·06	2·51	2·96
% *of Max.*	*5·0*	*6·4*	*7·3*	*9·2*	*10·6*	*4·6*	*8·4*

447. These average standard deviations expressed as percentages of maximum marks vary from 4·6 to 10·6. Naturally we expect these to be greater than the corresponding figures for the whole examination, because the random variations to a certain extent cancel out in the addition of marks to get the grand total.

448. With both Boards there is less precise marking in the case of Paper II, Qn. 2, than in the case of Qn. 1. The largest figure is that for "Style." We should anticipate the reverse of exact marking when an element "Style" is being considered.

449. We may profitably consider again Table 24 (para. 82), this time with the object of indicating one source from which the random variations in total marks are presumably derived. We have already pointed out that the adjusted average total marks of Examiners C and E are very nearly the same. Actually those of C, D and E are very near, both adjusted and unadjusted.

450. The Table below is derived from Table 24 (para. 82), and shows by how much the average marks for C, D, E for each

TABLE 125

Examiners C, D, E. Averages above (+) or below (−) the general average.

Question	Paper I				Paper II			
	Dictation	2a	2b	2c	1	2	Unadjusted Average Per cent.	Adjusted Average Per cent.
Maximum	11	15	20	15	30	20		
Examiner C	+0·21	−0·20	−0·28	+0·41	+0·21	−0·46	48·3	47·2
„ D	−0·26	+0·40	+0·50	+0·25	−0·77	+0·98	49·3	47·8
„ E	+0·16	−0·62	−0·64	−0·19	+0·03	+0·10	47·3	46·8

question are in excess or defect of the average of the examiners' averages (row *c*).

451. Here we see that C's averages are above the average in respect of three questions, and below in respect of three. D's are above in respect of four questions and below in respect of two, and E's are above in respect of three questions and below in respect of three. But there is not one question for which all three examiners are above (or below) the general average.

452. Examiners C and D are both above the average in the case of Qn. 2c of Paper I, but are at variance in this respect in all other questions. D and E are in agreement in the case of Qn. 2 of Paper II, but again are at variance in respect of all the others. Examiners C and E show different signs in respect of Qns. 2c of Paper I and 2 of Paper II. Moreover the amounts by which they differ from the general average are not the same, so that in some questions there are relatively large differences of standard.

453. Thus Examiner C's average is greater than E's average for Qn. 2c of Paper I by 0·62 marks (out of 15) and lower for Qn. 2 of Paper II by 0·56 marks (out of 20).

The difference between C and D in the case of Qns. 1 and 2 of Paper II is worthy of attention. In the first question C's average is greater than D's by 0·98 marks (out of 30) and in the second less by 1·44 marks (out of 20).

454. We have previously (para. 387) put forward the suggestion that such differences of standard in the marking of individual questions as have been here noted tend to introduce into aggregate marks for a paper the kind of discrepancies which we have described as due to random variations, that is, that some of the random element in aggregate marks is due to variations of standard from one question to another, some questions being marked rather generously by one examiner and more severely by another, and vice versa.

Section 3.—*School Certificate Chemistry* (see paras. 94–120)

455. The ideal marks were obtained for each Board. They are shown below, together with the differences between them, and the respective places in the orders of merit, and their differences. The orders of merit are also shown, together with the division lines between the various classes of Distinction, Credit, Pass, Failure.

TABLE 126

Cand. No.	Ideal Marks: Board I	Board II	Difference (I - II)	Class Award: Board I	Board II	Order: Board I	Board II	Difference (I - II)	Candidates Arranged in Order of Merit: Board I Candidate No.	Board II Candidate No.
1	48	53	− 5	C	C	15	17½	−2½	D 30	30 D
2	74	78	− 4	D	D	3	3	0	27	22
3	67	72	− 5	C	C	5	8½	−3½	2	2
4	43	46	− 3	P	P	19½	21	−1½	3	27
5	29	30	− 1	F	F	25	27	−2	8	8
6	17	21	− 4	F	F	28	28	0	15	15
7	41	40	+ 1	P	F	22	24	−2	22	21
8	67	73	− 6	C	D	5	6	− 1	9	3
9	64	72	− 8	C	C	8	8½	−0½	21	9
10	12	9	+ 3	F	F	29	30	−1	23	23
11	23	33	−10	F	F	27	25	+2	24	24
12	47	50	− 3	P	P	16	20	−4	C 29	29 C
13	43	55	−12	P	C	19½	15	+4½	14	25
14	50	54	− 4	C	C	13	16	−3	16	16
15	67	73	− 6	C	D	5	6	−1	1	13
16	49	57	− 8	C	C	14	14	0	12	14
17	34	44	−10	P	P	24	22	+2	25	1
18	25	32	− 7	F	F	26	26	0	4	28
19	11	16	− 5	F	F	30	29	+1	13	26
20	35	42	− 7	P	F	23	23	0	P 26	12 P
21	61	73	−12	C	D	9	6	+3	28	4
22	65	80	−15	C	D	7	2	+5	7	17
23	58	66	− 8	C	C	10	10	0	20	20
24	55	65	−10	C	C	11	11	0	17	7
25	46	60	−14	P	C	17	13	+4	5	11
26	43	52	− 9	P	P	19½	19	+0½	18	18
27	75	76	− 1	D	D	2	4	−2	11	5
28	43	53	−10	P	C	19½	17½	+2	F 6	6 F
29	53	64	−11	C	C	12	12	0	10	19
30	85	89	− 4	D	D	1	1	0	19	10
Average	47·7	54·3								

Distribution of awards, from ideal marks ; adopting, as in the case of French, for each Board its own limits between the different categories.

	Board *I*	*II*
Fail	6	8
Pass	9	4
Credit	12	11
Distinction	3	7
Total	30	30

456. The averages of the ideal marks differ by 7 marks. The examiners of Board II awarded higher marks on the average than those of Board I. The differences between the two sets of ideal marks are not the same in all cases, consequently the order of merit of the candidates given by the ideal marks of Board I is different from that given by the ideal marks of Board II. The resulting classification is also different, for although the ideal marks of Board II are on the whole higher than those of Board I by about 7 marks, and the limits of the various classes are different for the two Boards, the differences between these limits are not exactly 7 marks, thus :—

	Board I	*Board II*	*Difference*
Pass	34	43	9
Credit	48	53	5
Distinction	72	73	1

The combination of the different ideal marks with the different limits for the categories of Distinction, Credit, etc., has the effect of " spreading " the awards of the candidates more for Board II than for Board I.

457. The last two columns in Table 126 in para. 455 show the candidates placed in order according to the two sets of ideal marks, together with the class divisions.

It is apparent that if each Board had arranged for its own examiners to act finally in concert to issue agreed results on the bases of the ideal marks, the lists of the two Boards so settled would have been different.

458. We may point out, however, that if Board II's limiting marks for the classes had been

Pass	40
Credit	54
Distinction	78

in each case 6 marks greater than the corresponding limits for Board I, the numbers in the various classes would have been

Fail	6
Pass	8
Credit	13
Distinction	3

agreeing very nearly with those of Board I, and the candidates placed in the various classes would have been the same except for the following : Candidate No. 22 receives Distinction instead of Candidate No. 27, who obtains a Credit ; Candidates Nos. 25, 13 receive a Credit instead of Candidate No. 1, who obtains a Pass.

459. On the other hand, if Board I's limiting marks for the various classes had been

Pass	36
Credit	46
Distinction	66

in each case 7 marks lower than the corresponding limits of Board II, the numbers in the various classes would have been

Fail	8
Pass	5
Credit	11
Distinction	6

agreeing very nearly with those of Board II, and the candidates placed in the various classes would have been the same with the following exceptions: Candidates Nos. 22 and 21 now receive a Credit instead of Distinction, and Candidate No. 3 obtains a Distinction instead of a Credit; Candidates Nos. 13 and 28 now receive a Pass instead of a Credit, and Candidate No. 12 obtains a Credit instead of a Pass; Candidate No. 7 receives a Pass instead of a Failure and Candidate No. 17 is put in the Failure class instead of obtaining a Pass.

460. It is apparent that the most substantial difference between the two Boards' ideal classifications is due to the fact that the class limiting marks differ by varying amounts, the difference between the limiting marks for Distinction being only 1 mark, whereas the difference between the limiting marks for Pass is 9 marks.

461. The constant differences between the marks of individual examiners and the ideal are not very great owing to the method of arranging for the marking to follow well defined instructions. The extreme case is that of J, whose average is 51·6, 2·7 below the ideal average 54·3.

462. The standard deviations of the random marks are shown below.

TABLE 127

Standard Deviations of Random Variations

	Board I		*Board II*
Examiner A	2·6	Examiner G	5·5
,, B	4·0	,, H	3·1
,, C	2·6	,, J	4·7
,, D	4·2	,, K	3·6
,, E	4·0	,, L	2·7
,, F	3·6	,, M	2·8
Standard Deviation of ideal marks	18·6		19·8

The random element is not very pronounced, ranging from about $2\frac{1}{2}$ to $5\frac{1}{2}$ marks in 100. It is higher than in the corresponding

French examination, where the random marks had standard deviations ranging from 1·8 to 3·8. We may note that G's random marks on the average are about twice as large as those of L or M.

463. It is interesting to observe the effect of the random marks in the case of Examiner C of Board I. In his case the average mark, 47·4, is practically the same as the average of the ideal marks (47·7). The Table below shows the ideal marks (with awards), C's marks (with awards), and the differences between them.

TABLE 128

Candidate	Marks		Difference	Awards	
	Ideal	Examiner C	C from Ideal	Ideal	Examiner C
1	48	51	+3	C	C
2	74	71	−3	D	C
3	67	66	−1	C	C
4	43	41	−2	P	P
5	29	25	−4	F	F
6	17	16	−1	F	F
7	41	42	+1	P	P
8	67	74	+7	C	D
9	64	62	−2	C	C
10	12	16	+4	F	F
11	23	21	−2	F	F
12	47	47	0	P	P
13	43	46	+3	P	P
14	50	53	+3	C	C
15	67	71	+4	C	C
16	49	46	−3	C	P
17	34	34	0	P	P
18	25	20	−5	F	F
19	11	9	−2	F	F
20	35	34	−1	P	P
21	61	61	0	C	C
22	65	63	−2	C	C
23	58	58	0	C	C
24	55	56	+1	C	C
25	46	49	+3	P	C
26	43	41	−2	P	P
27	75	75	0	D	D
28	43	39	−4	P	P
29	53	51	−2	C	C
30	85	85	0	D	D
Average	47·7	47·4			

DISTRIBUTION OF AWARDS

	Ideal	Examiner C
Fail	6	6
Pass	9	9
Credit	12	12
Distinction	3	3

464. The numbers in the various classes agree as between the ideal and Examiner C, but individual results differ. The differences are shown below :—

Candidate	Ideal	Examiner C	Difference between C's mark and ideal (Random Variations)
2	D	C	− 3
8	C	D	+ 7
16	C	P	− 3
25	P	C	+ 3

Although the random variations are not great in the case of this examiner, in these four cases the variations are sufficiently large to mean a change in class.

465. Turning now to the individual questions, let us consider whether the random element is more pronounced in the case of one question in the paper than another. The Table below gives the averages of the examiners' standard deviations.

TABLE 129

Question	1	2	3	4	5	6	7	8
Number of Candidates	29	24	27	24	23	23	14	13
Board I								
Maximum Marks	16	16	16	16	16	17	16	16
Average Standard Deviation of Random Variations	0·58	1·52	1·30	1·40	1·12	1·33	1·14	1·35
Percentages of Maximum	*3·6*	*9·5*	*8·1*	*8·7*	*7·0*	*7·8*	*7·1*	*8·4*
Board II								
Maximum Marks	17	17	17	17	17	17	17	17
Average Standard Deviations of Random Variations	1·44	1·39	1·29	1·83	1·10	2·03	1·11	1·07
Percentages of Maximum	*8·5*	*8·2*	*7·6*	*10·8*	*6·5*	*11·9*	*6·5*	*6·3*

466. The random element on the whole appears to be present in the marking of each question to approximately the same extent. There are some slight differences, thus Qn. 4 has one of the highest figures in this Table in the case of both Boards, and Qns. 5 and 7 have two of the lowest figures. Answers to Qn. 4 appear to be less susceptible to precise marking than answers to Qns. 5 and 7.

The size of the percentage figures in this Table is roughly the same as in the case of French (see para. 446 above). Thus we may say that answers to French questions and Chemistry questions in the School Certificate examinations are marked with practically the same degree of precision. We may point out however that as Qn. 1 of Paper II in the French examination, which carried more than a quarter of the total marks, led to answers which were susceptible of considerable precision of marking, lower figures, 4·9 (Board I) and 4·6 (Board II), than those in Chemistry (except Qn. 1, Board I) being obtained for the average Standard Deviation of random variations, the analysis of the total marks shows that the marking of the French scripts on the whole was more precise than in the case of Chemistry.

467. An interesting sidelight on the presence of random marks in aggregates is furnished by a consideration of Table 34 (para. 116). The following figures extracted from that Table give the averages, for each question, of Examiners D and E and the differences between them.

Question	1	2	3	4	5	6	7	8	Average Whole Paper
Examiner D	8·24	5·75	6·93	6·00	8·26	9·26	8·64	10·46	45·9
„ „ E	8·24	6·33	7·41	5·21	8·43	8·04	8·64	9·69	45·5
Difference	0	−0·58	−0·48	+0·79	−0·17	+1·22	0	+0·77	+0·4

Roughly, the average marks awarded by these two examiners are the same for the total; but for individual questions, sometimes D marks more generously than E, sometimes less.

468. Such differences in standard of marking, from question to question, have the effect of introducing into aggregate marks some of the random element which has previously been the subject of discussion.

Section 4.—*School Certificate English* (see paras. 121–134)

469. The main conclusion is that, according to the ideal marks obtained, the following should have been the classification :

	Fail	Pass	Credit	Special Credit	Total
Candidates	1	20	27	0	48

Thus the candidates are all of moderate ability, the lowest ideal mark was 32, the highest was 58, the average was 45·9 and the standard deviation 6·04. These are all referred to a maximum of 100 marks.

470. In the Table below are shown the differences between each examiner's standard of marking and the ideal standard, and the sizes of the random variations are indicated by the appropriate standard deviations.

Examiner	Difference of Standards of Marking from Ideal	Standard Deviations of Random Variations
A	+2·5	4·12
B	+7·7	4·56
C	−6·2	3·27
D	+4·7	3·84
E	−1·2	3·12
F	+2·2	3·00
G	−7·3	4·27

471. Broadly, then, the best examiners introduce into their marking random errors of which the standard deviation is 3 marks out of 100, and the worst, random errors of which the standard deviation is 4½ marks. These appear to be small, but their effect is magnified when they are considered in relation to the natural variation[1] amongst the candidates, which in this case is only 6 marks out of 100, since, as is pointed out in para. 469, the candidates are all of moderate ability. Thus the extent of the examiners' variations introduced is from 50 to 75 per cent. of the real variation amongst the candidates, and this, of course, means that considerable changes are made in the grading of these candidates, who possess really very much the same ability in answering this paper. Further, the effect of the examiners' different standards is all-important, as the divergences here are very large compared with the variation in the group as a whole. B and G differ in their standards by 15 marks, which is more than half the whole range of variation in the ideal marks, 26 (32 to 58). Naturally, then, the awards made by these two examiners differ fundamentally, as was seen in the Table in para. 125.

472. The best examiners therefore might be as much as 6 marks (twice the standard deviation) wrong in their award of marks, whereas the worst examiner might be as much as 9 marks wrong in his estimates. To be on the safe side therefore in awarding Classes, Credits, Distinctions or other awards as a result of an examination of this nature, a good examiner should carefully scrutinise the marks of candidates within 6 marks of the borderline, and an examiner who is not so good should scrutinise those scripts which are 8 marks from the border. This means that if

[1] See footnote to para. 484.

there are 1,000 candidates in an examination with marks distributed in the following manner :—

Marks	Under 20	20–	30–	40–	50–	60–	70–	80 and over	Total
Candidates	50	100	150	200	200	150	100	50	1,000

and 50 per cent. is required for a Pass, then there are about 200 scripts which have to be carefully scrutinised by a good examiner, and perhaps 300 by a less competent examiner. But if the limit for a Pass is 40 per cent. of the marks and for Credit, 60 per cent., then a good examiner would have to scrutinise carefully about 350 scripts and a less good examiner about 550 scripts.

473. Let us now obtain from the figures in para. 470 the limits by which an ideal figure might be displaced by the examiners. The following Table shows these limits, assuming the range of variation to be roughly ± twice the standard deviation.

Examiner	Difference of Standards of Marking from Ideal	Standard Deviations of Random Variations	Range of possible Variation of a Mark per Script
A	+2·5	4·12	+2·5 ±8·2 i.e. − 6 to 11 marks
B	+7·7	4·56	+7·7 ±9·1 „ − 1 to 17 „
C	−6·2	3·27	−6·2 ±6·5 „ −13 to 0 „
D	+4·7	3·84	+4·7 ±7·7 „ − 3 to 12 „
E	−1·2	3·12	−1·2 ±6·2 „ − 7 to 5 „
F	+2·2	3·00	+2·2 ±6·0 „ − 4 to 8 „
G	−7·3	4·27	−7·3 ±8·5 „ −16 to 1 „

474. The final column of this Table shows the limits of the "errors" which the examiners might make in awarding marks for a given script. The Table below will illustrate the possible marks which might be awarded in specific cases :—

	Ideal Mark				
	34	40	46	53	60
Examiner	Limits of possible Variation of Marking				
A	28–45	34–51	40–57	47–64	54–71
B	35–51	41–57	47–63	54–70	61–77
C	21–34	27–40	33–46	40–53	47–60
D	31–46	37–52	43–58	50–65	57–72
E	27–39	33–45	39–51	46–58	53–65
F	30–42	36–48	42–54	49–61	56–68
G	18–35	24–41	30–47	37–54	44–61

475. We can appreciate the importance of this Table if we suppose for the sake of argument that 34 marks, 46 marks, 60 marks are the limits of grades. Suppose less than 34 marks means Failure, 34 marks and more but less than 46 means a Third Class, 46 marks and more but less than 60 means a Second Class, 60 marks and more means a First Class. We can construct a Table from the above showing the possibilities of awards by different examiners.

True Merit of Candidate in Classes		Marginal Third	Definitely Third	Marginal Second	Definitely Second	Marginal First
Ideal Mark		34	40	46	53	60
		Awards of Examiners				
A	Most likely	3rd	3rd	2nd	2nd	1st
	Possible	Fail	2nd	3rd	1st	2nd
B	Most likely	3rd	2nd	2nd	1st	1st
	Possible	2nd	3rd	1st	2nd	—
C	Most likely	Fail	3rd	3rd	2nd	2nd
	Possible	—	Fail	—	3rd	—
D	Most likely	3rd	3rd	2nd	2nd	1st
	Possible	Fail	2nd	3rd	1st	2nd
E	Most likely	3rd	3rd	3rd	2nd	2nd
	Possible	Fail	—	2nd	—	1st
F	Most likely	3rd	3rd	2nd	2nd	1st
	Possible	Fail	—	3rd	—	2nd
G	Most likely	Fail	Fail	3rd	2nd	2nd
	Possible	—	3rd	Fail	3rd	—

476. Thus a candidate who should really be considered as a possible Second Class (having 46 marks) might be " ploughed " by G, and if not " ploughed " would merely be awarded a Third Class. On the other hand, a candidate who should really be considered as barely attaining Third Class standard (having 34 marks) would possibly obtain a Second Class from B.

477. The numbers of candidates answering certain questions were as follows :—

	Paper I		*Paper II*				
			Question				
	Essay	Précis	1	4	5	10	13
Candidates	48	48	48	37	36	48	40

The analysis has been applied to the marks allotted to answers to these questions only, as the numbers involved in the other

questions of Paper II were smaller. For convenience when reference is made to these answers they will be noted as Essay, Précis, II 1, II 4, II 5, II 10, II 13. The maximum marks per question were not always the same; these will be indicated in the right place.

478. The differences between the examiners' standards of marking and the ideal are shown in the Table below, together with the averages of the ideal marks and the maximum marks per question.

Examiner	Essay	Précis	II 1	II 4	II 5	II 10	II 13	Total	Per cent. of Maximum
A	+2·1	+1·6	−0·9	+1·1	+0·9	0·0	+0·2	+5·0	+3·7
B	+3·0	+1·1	+2·1	+0·7	+1·1	+1·7	+0·9	+10·6	+7·9
C	−1·5	+0·3	−1·4	−1·7	−1·0	−0·5	−1·3	−7·1	−5·3
D	+1·9	−0·4	+0·5	+0·8	+1·8	+2·2	+0·7	+7·5	+5·6
E	+2·3	−0·6	−0·5	0·0	−0·2	−0·4	+0·2	+0·8	+0·6
F	+0·9	+1·2	+1·4	+0·9	+0·4	0·0	+0·2	+5·0	+3·7
G	−3·4	−2·1	−0·9	−0·6	−1·3	−0·7	+0·2	−8·8	−6·6
Average of Ideal Marks	13·6	6·9	11·5	7·7	7·3	8·7	6·1	61·8	46·1
Maximum	30	20	20	16	16	16	16	134	100

479. It will be observed that examiners are on the whole consistent from answer to answer. B is generous, C and G are severe. The total effect of this fundamental difference of standards is seen if we consider a whole paper made up of 7 questions and refer the total of these differences to the maximum marks. Out of 100 marks B would on the average tend to award 8 marks more than necessary, G would award 6½ too few. The difference between B and G is apparently fundamental. In an examination of this kind where a candidate should be allotted 40 marks by the ideal examiner, on the average B would give him 48 and G would award 33 or 34. Now 34 marks might be the requirement for recommendation for a Pass, 48 marks might be the minimum required for a Credit. G would be considering the placing of this candidate in one class, B would be considering putting him in a higher class. There is really a class interval between the average standard of marking of B and G. Normally, if B and G were examining together, they would presumably attempt to reconcile their differences of standards, but is such a reconciliation possible when there is such a wide divergence between them? Is it not more likely that the one would impose his standards on the other?

480. The figures in the last column in the Table in para. 478

should be compared with the corresponding figures of para. 470. Naturally, they are not the same, since not all the marks awarded have been submitted to analysis (see para. 477), but they are very much alike, as should be the case.

481. The Table below gives the standard deviations of the random variations, together with the standard deviations of the ideal marks.

	Essay	*Précis*	II 1	II 4	II 5	II 10	II 13
Standard Deviations of ideal marks	2·18	1·70	2·34	1·67	1·98	1·69	2·17
	Standard Deviations of Random Variations						
Examiner							
A	3·14	2·38	1·20	1·66	1·19	0·99	1·39
B	3·88	2·07	1·72	1·28	1·82	1·45	1·38
C	2·00	1·16	1·30	1·04	1·02	0·91	0·96
D	3·61	1·96	1·53	1·35	1·41	1·67	1·25
E	2·54	1·42	1·34	1·95	1·13	1·29	1·12
F	2·08	1·45	1·03	1·23	1·03	0·62	0·81
G	2·28	1·65	1·58	1·51	1·48	1·21	1·51
Maximum Mark	30	20	20	16	16	16	16

482. These figures are best appreciated if they are related to the maximum marks awarded per question. They are shown below therefore as percentages of these maxima :—

	Percentages						
	Essay	*Précis*	II 1	II 4	II 5	II 10	II 31
Standard Deviations of ideal marks	7·3	8·5	11·7	10·4	12·4	10·6	13·6
	Standard Deviations of Random Variations						
Examiner							
A	10·5	11·9	6·0	10·4	7·4	6·2	8·7
B	12·9	10·3	8·6	8·0	11·4	9·1	8·6
C	6·7	5·8	6·5	6·5	6·4	5·7	6·0
D	12·0	9·8	7·6	8·4	8·8	10·4	7·8
E	8·5	7·1	6·7	12·2	7·1	8·1	7·0
F	6·9	7·2	5·1	7·7	6·4	3·9	5·1
G	7·6	8·2	7·9	9·4	9·3	7·6	9·4
Average of Examiners' Variations	9·3	8·6	6·9	8·9	8·1	7·3	7·5

483. We may make several observations on this Table. In the first place, of the Standard Deviations of the ideal marks

expressed as percentages of the maximum marks, the least is that for the Essay question.

484. Secondly, comparing the average of examiners' variations with the Standard Deviations of the ideal marks, we note that the former are greater than the latter in the case of the Essay and Précis, and are less than the latter in the case of the other questions. On the whole, more precise marking is possible in dealing with questions of Paper II than in dealing with the Essay or Précis. The total variation in the marks awarded, consisting, as it does, of a combination of natural variation[1] amongst the candidates themselves and the examiner's variation, is due less to the random element than to the natural variation in the case of the marking of answers to precise questions, but is due more to the random element than the natural variation in the case of marking Essay and Précis questions.

485. We note that there is some consistency in the examiners' precision of marking from one piece of work to another. The Table below shows the examiners placed in the inverse order of the size of the standard deviations in the Table in para. 481.

Examiner	*Essay*	*Précis*	II 1	II 4	II 5	II 10	II 13	AllQuestions
A	5	7	2	6	4	3	6	4
B	7	6	7	3	7	6	5	7
C	1	1	3	1	1	2	2	1
D	6	5	5	4	5	7	4	6
E	4	2	4	7	3	5	3	3
F	2	3	1	2	2	1	1	2
G	3	4	6	5	6	4	7	5

C and F are the good examiners; B and D are not so good.

486. The importance of the examiner's random variations is of course reduced when we are dealing with the totals of marks for many questions, because an examiner may add two random marks when marking one answer, but would perhaps take off a mark when dealing with another answer, and so on. We can estimate the importance of this factor if we suppose that the examination consists of answers to all the questions which have furnished the material for this analysis (this was of course not quite the case; see para. 477). The standard deviations of the

[1] It is to be noted that the phrase "natural variation," as we have used it, means the variation of the ideal marks. This variation is not, however, independent of the choice of examiners or of the method of examination used. It is quite possible, for instance, that with more discriminating methods the "natural variation," as deduced from the marks, would have a higher value. It may, however, be noted that, in the comparison of the methods of marking Special Place English essays by impression and by detailed marking, the natural variation of the essays was found to be practically the same by the two methods (see para. 533 under the heading "Standard deviation of ideal marks," and also para. 438 *ante*).

examiner's variations for the whole of such a paper can be obtained by taking the square roots of the sums of the squares of the standard deviations for each question. Thus for Examiner A the new standard deviation would be $\sqrt{3{\cdot}14^2 + 2{\cdot}38^2 + 1{\cdot}20^2 + \ldots}$.

These standard deviations are shown below, together with the total marks for the whole paper, and as percentages of the latter.

Examiner	Standard Deviations	Per cent. of Maximum
A	4·90	3·66
B	5·59	4·16
C	3·30	2·46
D	5·23	3·90
E	4·27	3·18
F	3·33	2·48
G	4·32	3·22
Maximum Marks	134	100

487. The figures in the last column of this Table should be compared with the corresponding figures of the Table in para. 470. They are of course not the same, mainly because the figures in para. 470 are affected by the differences of standards of marking of the various pieces of work, and partly because other marks are also included in the original totals.

Section 5.—School Certificate Latin (see paras. 28–57)

488. The ideal marks obtained for the two groups of examiners are given below, together with the differences between the candidates' deviations from the averages.

Cand. No.	Ideal Marks Group I	Ideal Marks Group II	Deviations Group I	Deviations Group II	Differences
1	41	44	−1·00	−0·87	−0·13
2	44	45	+2·00	+0·13	+1·87
3	47	50	+5·00	+5·13	−0·13
4	42	43	0	−1·87	+1·87
5	44	46	+2·00	+1·13	+0·87
6	48	51	+6·00	+6·13	−0·13
7	46	50	+4·00	+5·13	−1·13
8	46	49	+4·00	+4·13	−0·13
9	37	41	−5·00	−3·87	−1·13
10	40	41	−2·00	−3·87	+1·87
11	39	44	−3·00	−0·87	−1·13
12	42	44	0	−0·87	+0·87
13	40	42	−2·00	−2·87	+0·87
14	40	45	−2·00	+0·13	−2·13
15	34	38	−8·00	−6·87	−1·13
Average	42·00	44·87			

489. We note that there is some difference between the two sets of ideal marks obtained from Group I and Group II of the examiners, and this may be due to these groups being composed of less and more generous markers, or the difference may be due to the fact that as far as one part of Paper I is concerned the examiners in Group II received less detailed instructions as to marking. This possibility must be examined.

490. But first it is interesting to compare more closely the two sets of ideal marks. If the difference between the average ideal marks is eliminated by comparing the deviations of each candidate's ideal marks from the average of the group, we can see better how closely the two groups are in agreement as to relative merits of the candidates.

Thus, when we eliminate the difference between the standards of the two groups, they are in agreement practically in 4 cases, differ by 1 mark in 7 cases, and by 2 marks in 4 cases. The respective orders of merit are :—

Candidates	
Group I	Group II
6	6
3	3
7	7
8	8
5	5
2	2
12	14
4	12
1	1
10	11
14	4
13	13
11	10
9	9
15	15

491. As each group of examiners ranks them in ten classes we can show the order of merit in this way :—

	Order of Merit		
Cand.	*Group I*	*Group II*	Difference
1	6	6	0
2	4	5	1
3	2	2	0
4	5	7	2
5	4	4	0
6	1	1	0
7	3	2	1

Cand.	Order of Merit *Group I*	*Group II*	Difference
8	3	3	0
9	9	9	0
10	7	9	2
11	8	6	2
12	5	6	1
13	7	8	1
14	7	5	2
15	10	10	0

There is therefore fair agreement between the two groups of examiners as to the respective merits of the candidates.

492. The differences between the examiners' standards of marking and the ideal are shown in the Table below.

Group I Examiners						*Group II* Examiners						
A	B	C	D	E	F	G	H	J	K	L	M	N
−3·40	+3·07	+7·80	−2·33	−1·67	+0·47	−1·20	−2·73	+9·60	−4·67	+6·07	−1·40	−0·2

493. On the average there is 11 marks difference between the standards of A and C, and 14 marks difference between those of J and K.

The Table below gives the standard deviations of the random variations.

STANDARD DEVIATIONS OF RANDOM VARIATIONS

Group I Examiner		*Group II* Examiner	
A	1·45	G	3·25
B	1·69	H	1·48
C	2·66	J	2·92
D	2·72	K	2·15
E	2·09	L	2·41
F	0·88	M	1·92
		N	2·67
Average	1·91		2·40
Standard Deviation of ideal marks	3·76		3·65

494. The random element in the marking is not very large. This may be attributed to the instructions given to the examiners. The standard deviations of the ideal marks are also small. This of course is due to the fact that the scripts used were chosen as of approximately the same level of excellence.

An examiner with a small standard deviation is more precise in maintaining his marking standard from script to script, one

Paper II. Max. 50.			Paper I. Max. 50.		
	Ideal Marks			*Ideal Marks*	
Cand.	*Group I*	*Group II*	Cand.	*Group I*	*Group II*
1	29	31	1	13	13
2	26	27	2	18	18
3	33	35	3	15	16
4	21	21	4	22	21
5	30	31	5	14	14
6	31	32	6	18	19
7	33	35	7	14	15
8	34	35	8	13	14
9	21	23	9	18	19
10	25	25	10	15	15
11	23	25	11	17	19
12	29	31	12	13	13
13	24	26	13	16	15
14	24	26	14	17	19
15	19	22	15	16	16
Average	26·80	28·33	Average	15·93	16·40

Paper II. Max. 50. Paper I. Max. 50.

AVERAGE OF EXAMINERS' MARKS

Group I		*Group II*		*Group I*		*Group II*	
A	24·87	G	25·53	A	13·73	G	18·13
B	27·47	H	27·67	B	17·60	H	14·47
C	32·67	J	33·00	C	17·13	J	21·47
D	24·80	K	25·67	D	14·87	K	14·53
E	25·27	L	31·40	E	15·07	L	19·53
F	26·40	M	27·40	F	16·07	M	16·07
		N	27·27			N	17·33

DIFFERENCE BETWEEN EXAMINERS' STANDARDS OF MARKING AND THE IDEAL.

Group I		*Group II*		*Group I*		*Group II*	
A	−1·93	G	−2·80	A	−2·20	G	+1·73
B	+0·67	H	−0·67	B	+1·67	H	−1·93
C	+5·87	J	+4·67	C	+1·20	J	+5·07
D	−2·00	K	−2·67	D	−1·07	K	−1·87
E	−1·53	L	+3·07	E	−0·87	L	+3·13
F	−0·40	M	−0·93	F	+0·13	M	−0·33
		N	−1·07			N	+0·93

STANDARD DEVIATIONS OF RANDOM VARIATIONS

Group I		*Group II*		*Group I*		*Group II*	
A	1·77	G	2·20	A	0·98	G	1·48
B	1·01	H	1·35	B	0·70	H	0·93
C	1·50	J	1·49	C	1·68	J	2·32
D	1·37	K	1·74	D	1·77	K	1·20
E	1·31	L	1·19	E	1·02	L	1·78
F	1·08	M	0·85	F	0·62	M	1·08
		N	2·49			N	1·81
Average	1·34		1·62		1·13		1·51

with a large standard deviation is less consistent. In the first Group, on the evidence before us, F is the best examiner in this respect, C and D are the worst. In Group II, Examiner H is the best, G and J are the worst. The standard deviations of the second Group on the whole are higher than those of the first Group. This greater precision of the first Group of examiners may partly be due to the more detailed instructions given to this group in respect of the marking of the unprepared translation of Paper I, or it may be merely an accidental difference between the two Groups. This point must now be looked into.

495. We can test the difference between the two groups by considering their marking of Paper II. In this case the instructions were identical. The previous analysis was repeated on the marks awarded for competence in this paper. The results are summarised on page 226, together with those for Paper I.

496. There is apparently some difference between the two groups of examiners. The average ideal mark of Group II is 1·53 higher than that of Group I in Paper II. The difference, 2·87, observed between the two ideal averages for the whole paper cannot therefore be attributed to the different instructions as to the marking of part of Paper I. In fact, roughly half of this 2·87 is accounted for in Paper II and the remainder, to be allocated to Paper I, might fairly be attributed to the accident that Group II chanced to consist of examiners on the whole more lenient than those in Group I. We may conclude that the different sets of instructions in Paper I were without influence on the result.

497. The figures showing the difference between the standards of the examiners and the ideal (in Paper II) tell nearly the same tale as before. In Group I, C is generous, A and D are more severe ; in Group II, J and L are generous, G and K are severe.

498. At this stage it is worth while making a comparison between the marking of Paper I and that of Paper II, remembering that Paper I consisted of questions on grammar, etc., and Paper II on Prescribed Books. In general the results show that examiners are consistently severe or generous with their marks on both papers. It is questionable whether the evidence is sufficient to show that examiners approximate more closely to the ideal when marking Paper I or Paper II.

499. The averages of the standard deviations of the examiners are less for Paper I than for Paper II in the case of both groups :—

	Paper I	Paper II
Group I	1·13	1·34
Group II	1·51	1·62

These figures suggest that Paper I is marked more precisely than Paper II. On the other hand, a glance at the figures for the individual examiners shows that more precise marking in Paper I compared with Paper II is achieved by Examiners A, B, E, F of Group I, and G, H, K and N of Group II, while the contrary is the case with Examiners C and D of Group I, and J, L, and M of Group II.

500. In conclusion, we may reasonably assert that the discrepancies between the examiners are mainly due to their different standards of marking rather than to their introduction of random variations in the marking.

Section 6.—*English Scholarship Essay* (see paras. 298–311)

501. The ideal marks are given in the Table below :—

Candidate	Ideal Marks	Candidate	Ideal Marks
1	36	26	61
2	48	27	55
3	64	28	47
4	33	29	54
5	53	30	42
6	49	31	54
7	50	32	51
8	64	33	56
9	46	34	56
10	65	35	66
11	62	36	57
12	56	37	62
13	52	38	55
14	55	39	50
15	61	40	55
16	65	41	57
17	57	42	50
18	58	43	56
19	47	44	53
20	30	45	59
21	57	46	40
22	70	47	37
23	48	48	50
24	44	49	58
25	57	50	18

The average of the ideal marks is 52·5, and their standard deviation 9·8.

502. The distribution of candidates into classes according to the ideal marks is as follows.

Class	Number of Candidates
I	1
II	35
III	12
IV	2
Total	50

This classification should be compared with Table 97, p. 144 above. The difference between the ideal classification and that of individual examiners, especially in respect of Class I, is striking.

503. The differences between the examiners' standards of marking and the ideal are given in the Table below. These differences are comparatively small.

Examiner				
A	B	C	D	E
−0·7	+0·1	+2·3	+1·4	−1·9

504. The standard deviations of the random variations are given below. These are by contrast rather large.

Examiner					Standard Deviation of ideal marks
A	B	C	D	E	
6·8	9·1	9·0	7·5	6·2	9·8

It is apparent that the large discrepancies already noted in Part I, paras. 302–308, are due more to the random variations introduced into their marking by the examiners than to differences of standards of marking.

505. There were four subjects for the essay, of which one was chosen by each candidate. It seems worth while to investigate the relationship between the size of the random variations and the subject of the essay.

Of the fifty candidates, ten, those numbered 1–10, wrote on one subject; eight, those numbered 11–18, wrote on another subject; eleven, those numbered 19–29, wrote on a third; and twenty-one, those numbered 30–50, wrote on a fourth. The random variations introduced by the examiners were considered in these four groups, and averages were obtained.

Subject	Candidates	Examiner A	B	C	D	E
1	(1–10)	+1·8	−2·9	−2·3	+0·7	+1·8
2	(11–18)	+6·4	−0·4	−0·2	−4·7	−0·4
3	(19–29)	−0·5	−5·4	+1·2	+2·4	+1·6
4	(30–50)	−2·2	+4·7	+1·2	+1·3	−1·3

506. The random variations introduced by Examiner A in marking scripts (11–18) are as follows: + 4, − 1, + 16, + 8, − 5, + 5, + 16, + 8; those of Examiner D marking the same

scripts are − 8, − 7, − 11, − 14, + 4, − 3, − 3, + 4. The averages of these are + 6·4 and − 4·7 respectively. It seems reasonable to suppose that it is worth while investigating the possibility that these mostly positive variations of Examiner A, and the negative variations of Examiner D, are connected with the examiner's reactions not merely to essays as such but to those particular essays on certain subjects. *A priori* we might suppose that Examiner A awarded extra marks to those candidates writing on Subject 2; that Examiner D took off marks; and that Examiner B gave low marks to those candidates taking Subject 3, and awarded high marks to those taking Subject 4.

507. This possibility was therefore explored by dividing the whole fifty scripts into the four groups determined by the subject of the essay. Each group was analysed separately in the same way as the original group was treated. The following results were obtained :—

Differences between Standards of Marking and Ideal

		Examiner				
		A	B	C	D	E
Subject	Candidates					
1	(1–10)	+0·9	−2·8	−0·2	+1·8	−0·1
2	(11–18)	+4·1	−1·6	+0·5	−5·0	−3·6
3	(19–29)	−1·5	−5·5	+3·1	+3·3	−0·5
4	(30–50)	−2·2	+4·7	+4·3	+3·3	−2·3

Standard Deviations of Random Variations

		Examiner				
		A	B	C	D	E
Subject	Candidates					
1	(1–10)	5·7	6·9	9·5	7·4	7·5
2	(11–18)	8·2	7·0	1·2	5·9	8·1
3	(19–29)	3·3	10·3	6·2	6·5	8·9
4	(30–50)	5·6	8·0	12·0	8·5	3·2
	Averages	5·7	8·0	7·2	7·1	6·9

508. There appears to be some justification for our suggestion, although it must be remembered that the first three groups only contain ten, eight, eleven candidates respectively, so that the results obtained cannot be considered as being very precise. But we note that the difference in standards of marking in some groups are quite large, e.g. in the group for Subject 3, Candidates (19–29), Examiner B's standard differed considerably on the whole from that of Examiner D.

509. Examiner A apparently awarded extra marks to candidates taking Subject 2; Examiner B preferred those taking Subject 4, and did not react favourably to Subject 3; Examiner C tended to mark highly on the whole, but preferred Subjects 3

and 4 ; Examiner D marked rather high, and preferred Subjects 3 and 4 and did not favour Subject 2 ; Examiner E marked on the low side generally, and definitely reacted unfavourably to Subject 2.

510. When we consider the random variations, again we find differences. Thus the random element is very small when Examiner E marks scripts on Subject 4, and when Examiner C marks those on Subject 2. On the other hand Examiner C appears to introduce a large element of randomness in his marking of Subject 4.

Further, on the average for all the examiners except E, the random variations introduced into the marking are seen to be smaller when the candidates are dealt with in these four groups than when they are considered as a whole. This suggests that the differences in standards of marking between subjects are real.

Section 7.—*History Honours* (see paras. 321–342)

511. The method of obtaining the size of random variations in an examiner's marks which is used when numerical marks are available is no longer possible when the results are given as literal marks. But by a modification of the method we can form estimates of the relative incidence of random variations in marking as between one examiner and another. If no element of randomness is present in the marks awarded by two examiners to a set of scripts the correlation between the orders of merit determined by the marks of the two examiners will be perfect. The entry of the random element disturbs the orders of merit, and the correlation coefficient becomes less than unity. If the random variations are not very large, the correlation coefficient will not differ from unity by a great deal, but if the random element in the marking is considerable, the coefficient of correlation may become zero or even be negative.

512. It is on this principle that we rely to obtain an estimate of the random element in marking in the present investigation. We can, from the literal marks given, obtain orders of merit for each examiner, and combining these in pairs we may get the appropriate correlation coefficients which will serve to indicate the presence of large or small random variations. A little manipulation of the correlation coefficients gives us the relationship between the standard deviation of the random variations and the standard deviation of the ideal marks for each examiner. We cannot estimate the standard deviation of the ideal marks, therefore we cannot determine absolutely the

TABLE 130

Correlation Coefficients

Paper I

				DK	·32	DO	·49	DP	·42	DQ	·53							
						KO	·64	KP	·31	KQ	·31							
								OP	·36	OQ	·37							
										PQ	·42							

Paper II

AB	·36	AC	−·02	AF	·16	AH	·19	AJ	·34	AK	−·10	AL	·44	AN	·26	AR	·35
		BC	·15	BF	·27	BH	·23	BJ	·04	BK	·19	BL	·53	BN	·47	BR	·36
				CF	·49	CH	·30	CJ	·44	CK	·38	CL	·12	CN	·29	CR	·01
						FH	·28	FJ	·28	FK	·56	FL	·31	FN	·45	FR	·11
								HJ	·09	HK	·23	HL	−·14	HN	−·03	HR	−·03
										JK	·25	JL	·32	JN	·45	JR	·18
												KL	−·10	KN	·51	KR	·35
														LN	·59	LR	·36
																NR	·50

Paper III

AB	·04	AF	·20	AH	−·02	AJ	·47	AK	·65	AL	·31	AN	·43	AQ	·29	AR	·23
		BF	·47	BH	·27	BJ	·33	BK	·15	BL	·73	BN	·51	BQ	·47	BR	·70
				FH	·29	FJ	·50	FK	·60	FL	·56	FN	·78	FQ	·75	FR	·48
						HJ	·36	HK	·04	HL	·44	HN	·21	HQ	·13	HR	·22
								JK	·34	JL	·73	JN	·68	JQ	·51	JR	·51
										KL	·38	KN	·49	KQ	·58	KR	·30
												LN	·63	LQ	·73	LR	·85
														NQ	·79	NR	·60
																QR	·66

Paper IV

AB	·43	AE	·49	AF	·29	AG	·37	AH	·26	AJ	·42	AL	·14	AM	·55	AQ	·33
		BE	·75	BF	·47	BG	·31	BH	·64	BJ	·17	BL	·64	BM	·69	BQ	·55
				EF	·68	EG	·24	EH	·75	EJ	·13	EL	·41	EM	·61	EQ	·70
						FG	−·05	FH	·79	FJ	·17	FL	·30	FM	·38	FQ	·68
								GH	·23	GJ	·50	GL	·64	GM	·44	GQ	·31
										HJ	·22	HL	·60	HM	·73	HQ	·83
												JL	·39	JM	·31	JQ	·20
														LM	·55	LQ	·61
																MQ	·57

standard deviations of the random variations, but we can compare one examiner with another in respect of random marking.

513. The Table on page 232 shows the correlation coefficients referred to above.

The highest value of the correlation coefficient is 0·85, between the marks of L and R in Paper III ; the lowest is —0·14 between H and L in Paper II. In the latter case the random element in the marking of H or L is so large as to affect the common element in the correlation of the natural differences between the candidates.

514. The results of the calculations to determine the relative size of the random element in the marking are shown below :—

TABLE 131

RATIO OF RANDOM VARIATIONS TO THE NATURAL VARIATION OF THE GROUP OF CANDIDATES

	Paper			
Examiner	I	II	III	IV
A		2·15	2·24	1·50
B		1·44	1·34	0·80
C		1·93		
D	1·03			
E				0·75
F		1·17	0·82	1·32
G				1·84
H		4·31	3·15	0·61
J		1·65	0·91	2·34
K	1·34	1·79	1·44	
L		1·61	0·43	1·00
M				0·71
N		0·74	0·58	
O	0·88			
P	1·46			
Q	1·24		0·69	0·74
R		1·88	0·87	
Average	1·24	1·72	0·99	0·90

If a figure in the above Table is very small, it means that the examiner introduces little personal error of a random character into his marks ; if the figure is unity, it means that the examiner's random error is of the same size as the amount of variation amongst the candidates due to their inherent difference of ability in this subject ; if a figure is as high as 4, it means that the random error swamps the natural variation of the candidates.

Some of the figures in the Table above appear to be very large indeed, e.g. H's figures in Papers II and III. N appears to be comparatively precise ; H is precise when marking Paper IV, but introduces large errors when marking Papers II and III ; A consistently shows lack of precision.

515. In order to make comparisons possible with the results

of the other investigations, let us assume that the standard deviations of the ideal marks are all 10 out of 100. This is the measure of variability in a group of candidates of moderate ability, which these appear to be. On this basis the average standard deviations of the random variations would be

Paper	I	II	III	IV
	12	17	10	9

516. From the average of these figures indicating random marking, it appears that Papers III and IV are easier to mark than the others, and that Paper II is the most difficult paper on which to get examiners to agree.

517. From the figures we have given we are justified in drawing as a general conclusion that even when examiners in History use literal marks instead of numerical marks, the same striking discrepancies enter into the results; there are still differences of standard, and there are still considerable variations of a random character introduced into the marking by the examiners.

Section 8.—*Mathematical Honours* (see paras. 312–320)

518. The Table below shows the original marks awarded by the six examiners, together with the approximations to the ideal marks. The maximum was 300.

Examiner	A	B	C	D	E	F	Ideal
Candidate							
1	209	185	223	235	225	212	215
2	200	205	180	193	205	208	198
3	201	208	172	198	197	179	193
4	175	193	172	177	212	189	186
5	81	94	81	100	123	145	103
6	200	217	203	205	207	187	204
7	119	140	137	157	134	150	140
8	167	201	187	198	190	190	190
9	147	155	127	139	140	147	143
10	203	220	203	192	205	208	205
11	85	66	79	78	108	65	80
12	133	122	140	128	127	133	130
13	224	228	239	253	222	241	235
14	215	226	228	223	234	217	224
15	224	245	255	262	216	245	242
16	95	120	136	143	135	127	127
17	165	161	171	168	178	177	170
18	287	294	290	308	300	303	297
19	123	101	66	100	114	102	101
20	154	125	118	122	163	175	141
21	117	102	120	131	136	113	120
22	89	73	75	81	75	87	80
23	271	278	277	287	273	282	278

519. The differences between the examiners' standards of marking and the ideal are shown below, together with the average of the ideal marks and the standard deviation of the ideal marks, and the standard deviations of the examiners' random variations.

Examiner	A	B	C	D	E	F	
Differences between standards of marking and ideal	−5·1	−1·9	−5·3	+3·3	+5·1	+3·5	Average 174
Standard Deviation of random variations	12·5	11·4	12·0	10·4	12·4	13·0	Standard Deviation of ideal marks 59·0

520. The amount by which the examiners' general level of marking differed from the ideal was comparatively slight, only 1·7–1·8 marks out of 100 at the most.

521. The extent of random variations in examiners' standards is between 3·5 and 4·3 marks per 100, which appears to be of considerable magnitude, especially in view of the generally accepted idea that marking of Mathematics papers is not liable to error.

522. We may consider the random variations introduced into the marking again from a slightly different point of view. If the ideal mark is Q, and random variations of a size indicated by a standard deviation of 12½ marks (say) are likely to be introduced, it is reasonable, from our knowledge of the occurrence of random events, to assume that the actual mark awarded would fairly certainly be between Q − 25 and Q + 25 (a range of twice the standard deviation on either side of Q). Thus any candidates within 25 marks of the borderline of a class should be reconsidered by an examiner, so that a fair decision may be made as to the appropriate class into which an individual should be placed.

523. If, for instance, 210 marks was the minimum requirement for a First Class, then all those candidates with marks between 185 and 235 should be considered as possibly coming in that category. Consequently Examiner A would carefully reconsider the scripts of Candidates Nos. 1, 2, 3, 6, 10, 13, 14 and 15, with the object of deciding which of them are really First Class. An examination of the marks awarded by the examiners, and the ideal marks, shows how necessary this reconsideration really is.

The following Table shows candidates' numbers on the First Class borderline, according to the original marks :—

Examiner					
A	B	C	D	E	F
Candidates					
1	1	1	1	1	1
2	2	6	2	2	2
3	3	8	3	3	4
6	4	10	6	4	6
10	6	14	8	6	8
13	8		10	8	10
14	10		14	10	14
15	13			13	
	14			14	
				15	

524. The classes which would be awarded by the examiners according to the original marks of all the candidates in the Table above are shown below :—

Candidate's Number	Examiner						Ideal
	A	B	C	D	E	F	
1	2	2	1	1	1	1	1
2	2	2	2	2	2	2	2
3	2	2	2	2	2	2	2
4	2	2	2	2	1	2	2
6	2	1	2	2	2	2	2
8	2	2	2	2	2	2	2
10	2	1	2	2	2	2	2
13	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1

The six examiners and the ideal grading are in agreement in six cases, and some difference exists in the case of the other four. (It is only fair to state that in this investigation the examiners were not actually requested to grade the candidates in classes.)

525. The common practice in examinations of this type is that the scripts are examined by two examiners, so that the revision discussed above actually does take place. In order to test the extent of the discrepancies still remaining after this revision has taken place, the original six examiners were formed into three pairs, and these pairs reconsidered the scripts and produced agreed marks.

526. The revised marks of the pairs A B, C D, E F were examined in the same way as the original marks, and ideal marks

were again obtained. The Table below shows these, together with the ideal marks obtained from the original six sets for comparison with the new ideal marks :—

Pair	AB	CD	EF	Ideal	Ideal (from Original Six)	Differences between ideal marks
Candidate						
1	198	230	219	213	215	+2
2	203	183	207	199	198	−1
3	203	186	190	194	193	−1
4	186	177	210	191	186	−5
5	86	96	128	101	103	+2
6	207	208	195	204	204	0
7	125	145	142	136	140	+4
8	188	194	190	190	190	0
9	151	138	144	145	143	−2
10	216	203	207	210	205	−5
11	76	87	88	83	80	−3
12	128	137	130	131	130	−1
13	220	246	239	233	235	+2
14	220	226	225	223	224	+1
15	239	260	241	245	242	−3
16	117	136	131	126	127	+1
17	163	171	178	170	170	0
18	290	300	302	296	297	+1
19	113	91	108	105	101	−4
20	132	123	169	141	141	0
21	110	120	122	116	120	+4
22	79	83	81	81	80	−1
23	279	282	278	280	278	−2

The two sets of ideal marks are reasonably close to each other, never differing by more than 5 marks. It must be emphasised that each of these sets of marks is merely an approximation to the unknown ideal set of marks (cf. para. 440).

527. The differences between the examiners' standards of marking and the ideal are shown below, together with the average and standard deviation of the ideal marks and the standard deviations of the random variations.

Pair	AB	CD	EF	
Differences between Standards of Marking and ideal	−3·7	+0·4	+4·8	Average 174·5
Standard Deviations of Random Variations	6·9	9·7	8·9	Standard Deviation of ideal marks 59·1

528. The pairing of examiners has successfully reduced the amount of random variation introduced by individual examiners, but naturally some variation still remains. The Table below shows the extent of the change which has taken place :—

	Random Variations					
Examiner	A	B	C	D	E	F
Standard Deviation	12·5	11·4	12·0	10·4	12·4	13·0

Pair	AB	CD	EF
Standard Deviation	6·9	9·7	8·9

529. Moreover, the difference between standards of marking and the ideal have been changed somewhat, the earlier and revised figures being :—

	Differences between Standards of Marking and ideal					
Examiner	A	B	C	D	E	F
	−5·1	−1·9	−5·3	+3·3	+5·1	+3·5

Pair	AB	CD	EF
	−3·7	+0·4	+4·8

530. Thus, using examiners in pairs is successful both in obtaining a better general standard of marking and in reducing random variations due to individual examiners' personal idiosyncrasies. But even now, when classes are being discussed, the borderline cases should again receive special consideration before it is decided whether a candidate shall be placed in one grade or another. Thus the pair A B should reconsider candidates within 14 marks of the borderline mark, pair C D those within 20 marks of the borderline, and pair E F those within 18 marks of the borderline. If we illustrate again by considering the first-class borderline (210 marks), the following candidates should receive this special consideration.

Pair		
AB	CD	EF
Candidates		
1	1	1
2	6	2
3	8	4
6	10	6
10	14	10
13		14
14		

531. The number of such candidates is reduced ; and the classes into which they would be placed by the pair of examiners before this reconsideration are as follows.

Candidate's Number	AB	CD	EF	Ideal
1	2	1	1	1
2	2	2	2	2
3	2	2	2	2
4	2	2	1	2
6	2	2	2	2
8	2	2	2	2
10	1	2	2	1
13	1	1	1	1
14	1	1	1	1

The pairs and the ideal agree in six out of the nine cases, and there is disagreement in the other three. Again it is only fair to state that in this investigation the examiners were not asked to place the candidates in classes.

Section 9.—Essay Scripts at the Special Place Examination (II)
(see paras 246–297)

532. The material furnished in this investigation is very suitable for analysis by our methods. The major question at issue is this :—Is marking by Impression or by Details the more precise ? This question can be answered quite clearly.

533. As we have seen, the same ten examiners marked two sets of 75 essays, the one set by impression and the other by details. We have the following results. Table 132 below shows the differences between the standards of marking of the different examiners and the ideal, the averages and standard deviations of the ideal marks, and the standard deviations of the random variations.

TABLE 132

	Differences between Standards of Marking and Ideal		Standard Deviations of random variations	
Examiner	By Impression	By Details	By Impression	By Details
A	+2·0	+4·9	10·0	7·7
B	−3·3	−1·1	9·0	11·0
C	+12·4	+6·6	9·0	7·9
E	−15·2	+3·1	9·8	10·0
G	−2·4	+2·8	11·5	6·0
K	+0·5	−6·4	6·6	8·2
L	+4·2	−2·2	6·3	7·2
M	−7·0	−5·2	7·3	6·6
N	−0·8	+0·2	7·7	7·9
P	−5·3	−1·2	7·0	6·3
Average Difference	5·3	3·4	Average 8·4	7·9
Average of ideal marks	47·0	55·7		
Standard Deviation of ideal marks	14·4	13·9		

534. The differences between the standards of marking and the ideal are on the whole (taking all the examiners into consideration) less with detailed marking than with impression marking. But there is no sensible difference between the size of the random variations introduced into the two methods of marking, the average standard deviation being 8·4 by impression, and 7·9 by details.

535. Individual examiners differ in this respect; five (A, C, G, M, P) have standard deviations of random variations which are greater with marking by impression than with detailed marking, and the other five (B, E, K, L, N) show the opposite. In only one case, that of Examiner G, is the difference really large. This Examiner G is the only one who could be confidently regarded as exhibiting in his marks a precision which differs according to the method of marking; and in his case his marks are more precise when marking by details than when marking by impression.

536. The evidence of this experiment shows therefore that on the whole no greater precision of marking is obtained by details rather than by impression. The greater agreement between the marks in the former case is due to the fact that in detailed marking there is not so much difference between standards of marking. But as these differences between standards of marking can be eliminated by simple manipulations of the marks in an office, nothing apparently is gained by marking by details rather than by impression.

Section 10.—*Special Place Examination* (*I*) (see paras. 135–245)

537. In Part I, when the results of the investigation into the discrepancies observed in the marking of Arithmetic and English at the Special Place Examination were analysed, details were given of the marking of the various parts of one question, i.e. Qn. 1 of English, Part B (see paras. 214 *et seq.*). In particular we showed there a Table giving the relative frequency of disagreement of the examiners from the majority when marking the various parts of the question (see Table 73, para. 231).

538. The final figures from this Table will serve to indicate which examiners differ most from their colleagues and which differ least, and they might therefore be used to indicate which examiners are better than the others.

The figures to which we refer are :—

Examiner									
A	B	C	D	E	F	G	H	J	K
8·3	11·3	8·0	22·5	9·4	5·7	8·5	6·0	9·8	10·4

When we were considering the details of the question we saw that Examiner D was most often a disturber of agreement ; the figure corresponding to D in this Table is the largest. We conclude from this Table that F and H are the best examiners, since they are most often in agreement with the majority. These figures can be used to indicate the relative importance of the examiners, as far as this individual question is concerned.

539. If we submit the marks obtained by the 150 candidates from the ten examiners for this question to the same analysis as before, with the object of finding the extent of the random variations, we shall get results which may be compared with the foregoing, obtained from a direct attack on the actual known details of the marking, whereas the results we shall now give are obtained after analysis based on speculation as to the detailed method of allotting marks by examiners.

540. The standard deviations of the random variations are as follows :—

Examiner									
A	B	C	D	E	F	G	H	J	K
0·76	1·15	0·79	1·73	1·14	0·93	1·05	0·94	1·08	1·06

(the maximum mark for the question being 14).

The examiner with the largest figure in this series is D, again indicated as the worst examiner ; the best examiners are A and C according to these figures.

We do not expect that these two sets of figures will be exactly similar, because those in para. 538 above were obtained after a consideration of the details of a large percentage of the 150 candidates, but not all, whereas the figures just quoted are obtained from the analysis of all the 150 sets of marks. But there is a great resemblance between the two sets of indicators, the correlation between them being 0·91.

541. The fact that these two sets of figures give roughly the same indications as to good and poor examiners can be seen easily, if we put the examiners in order according to the two sets of figures.

	Examiner									
	A	B	C	D	E	F	G	H	J	K
Order of figures in para. 538	4	9	3	10	6	1	5	2	7	8
Order of figures in para. 540	1	9	2	10	8	3	5	4	7	6

(The better examiners being placed higher than the worse.)

542. Both sets of figures indicate that Examiners A, C, F, G, H are the better examiners, and B, D, E, J, K are the worse.

Section 11.—Summary of the Results

543. We may make comparisons between the different examinations by comparing the standard deviations of the random variations which are present in the marks awarded to answers to whole papers. A summary of the results given in the foregoing is shown in Table 133 below.

TABLE 133

STANDARD DEVIATIONS OF RANDOM VARIATIONS

(Max. in each case 100)

School Certificate

Exr.	History Investigation (1)	(2)	French Exr.	Bd. I	Exr.	Bd. II	Chemistry Exr.	Bd. I	Exr.	Bd. II	English Exr.	Latin Group Exr.	Bd. I	Exr.	Group Bd. II
A	4·9	—	A	3·8	G	1·8	A	2·6	G	5·5	A 4·1	A	1·4	G	3·2
B	3·9	5·5	B	2·5	H	3·7	B	4·0	H	3·1	B 4·6	B	1·7	H	1·5
C	7·0	7·0	C	2·5	J	2·7	C	2·6	J	4·7	C 3·3	C	2·7	J	2·9
D	4·4	5·2	D	2·7	K	3·1	D	4·2	K	3·6	D 3·8	D	2·7	K	2·1
E	7·2	8·0	E	2·4	L	2·1	E	4·0	L	2·7	E 3·1	E	2·1	L	2·4
F	7·3	5·0	F	3·3	M	2·5	F	3·6	M	2·8	F 3·0	F	0·9	M	1·9
G	7·1	8·0									G 4·3			N	2·7
H	2·3	4·2													
J	7·1	6·3													
K	4·6	5·4													
L	3·6	4·9													
M	3·2	3·1													
N	5·1	3·9													
P	5·8	5·6													
Q	4·6	6·1													
Average	5·2	5·6		2·9		2·7		3·5		3·7	3·6		1·9		2·4
S. D. of Ideal Marks	5·9	5·5		15·5		16·9		18·6		19·8	6·0		3·8		3·7

Exr.	English Schol. Essay	I	History Honours Paper II	III	IV	Exr.	Mathematical Honours Pairs		Special Place English Essay Exr. Impn.	Detailed
A	6·8					A	4·2	(AB) 2·3	A 10·0	7·7
B	9·1					B	3·8	(CD) 3·2	B 9·0	11·0
C	9·0					C	4·0	(EF) 3·0	C 9·0	7·9
D	7·5					D	3·5		E 9·8	10·0
E	6·2					E	4·1		G 11·5	6·0
						F	4·3		K 6·6	8·2
									L 6·3	7·2
									M 7·3	6·6
									N 7·7	7·9
									P 7·0	6·3
Average	7·7	12	17	10	9		4·0	2·9	8·4	7·9
S. D. of Ideal Marks	9·8	10	10	10	10		19·7	19·7	14·4	13·9

544. The average standard deviations may be arranged in this way :—

School Certificate Latin	1·9, 2·4
School Certificate French	2·9, 2·7
School Certificate English	3·6
School Certificate Chemistry	3·5, 3·7
Mathematical Honours	4·0, (Pairs 2·9)
School Certificate History	5·2, 5·6
English Scholarship Essay	7·7
English Special Place	8·4, 7·9
History Honours	12, 17, 10, 9

545. On the evidence, most precision is possible in the case of School Certificate Latin and French, where detailed instructions are possible and where in many cases an examiner has merely to compare a piece of work with a model. Least precision is possible in Essay type examinations. The positions occupied by School Certificate Chemistry and Mathematical Honours in this arrangement are worthy of attention. The general idea that mathematics and science subjects can be marked with greater precision than humanistic subjects is apparently not founded on a sound basis.

MEMORANDA

MEMORANDUM I

THE ANALYSIS OF EXAMINATION MARKS

BY

CYRIL BURT

Section 1.—*Introduction*

546. During the last half-century psychologists have been attempting to devise accurate and quantitative methods for standardizing the measurement of mental capacities among school children; and more recently these methods have been applied with considerable success to the standardization of tests for educational attainments in the more elementary subjects of the curriculum. It is natural to suggest that the same methods might be extended to the investigation, and possible refinement, of the means adopted for measuring or marking proficiency in higher subjects—to scholarship examinations or examinations for University degrees. It seems, for example, urgently desirable that every examiner should be able to say what degree of accuracy is obtainable in the examinations carried out in his own special branch, and particularly what degree of accuracy he himself and his colleagues are personally achieving under existing arrangements. I am far from claiming that a statistical analysis of marks is the only means of attacking the problem, much less that arithmetical devices can be applied to mental characteristics with the same rigorous exactitude that may be attainable in dynamics or astronomy. But it can hardly be disputed that some quantitative criterion, however rough or tentative, would at least afford a more precise and a more objective answer to questions such as I have instanced than a bare statement of general impressions or of personal experience.

In the hope, therefore, that examiners may feel urged to make some systematic analysis of their own work, and, in fact, to examine themselves, I have endeavoured to describe and

demonstrate what I take to be the best procedure at the moment available. Those who think of attempting such investigations and of applying the formulæ proposed will naturally wish to know what are the assumptions and the arguments on which the methods are based; hence I have sought to set out, however imperfectly, the principles by which they have been deduced. To non-mathematicians, who may feel alarmed by the array of symbols scattered over the following pages, I may point out that the most important calculation of all involves, in its simplest form, nothing more than the power to rank candidates in an order of merit, subtract the figures obtained for each of the candidates, and add up the rank-differences to find the total amount of discrepancy.[1]

Section II.—Preliminary Assumptions

547. A mark is a mental measurement—an approximate estimate on an arbitrary scale of an individual's capacities or attainments. What attainments and capacities are to be measured is presumably specified or implied in the syllabus of the examination. As with most forms of measurement, the observations are indirect; usually we are marking a script rather than a candidate. Moreover, mental measurements, far more than physical, include a large admixture of error. The examiner, as it were, is subject to the pull of various influences: the true value of the script is only one. Our primary object is to examine the nature, sources, and magnitude of these errors.

Two lines of approach are possible. To the scientific worker, trained in the technique of quantitative estimation, the most natural will be to resolve the concrete mark or measurement into its presumable components, so as to isolate the true value from the rest. To this end he will apply the technical devices, which to him are quite familiar, but will seem somewhat formidable to the plain man. For the latter the more intelligible

[1] See page 271, para. 567. The original draft of my Memorandum, based on earlier notes, was drawn up after the Conference at Folkestone, in July, 1935, when the simpler approximations used by Dr. E. C. Rhodes were subjected to some criticism. The main result of my own investigation is to show that the more exact and elaborate methods, such as would be used for psychological tests with large samples of school children, are hardly worth while with groups so small as we are dealing with here. At the same time, it would seem a sound policy to consider first the possible results of the more exact methods before deciding to accept those of the more approximate methods. As the foregoing chapters were already completed, my memorandum had to be revised under great pressure of time. I have, however, to thank Dr. Rhodes who very kindly read through my original draft, and was good enough to criticize its ambiguities and inexactitudes. I have also to acknowledge the help I have received from Mr. A. R. Kelly and from my wife, who have been good enough to assist me by reading through my manuscript and proofs, working out the illustrative calculations, and suggesting many welcome improvements in the text.

approach will be to start with the obvious influences that are bound to affect a human examiner, and see how these are compounded to give the resultant mark.

548. Let us glance first at the former alternative, the method of resolution or analysis. The purpose of an examination is to estimate each candidate's true mark. In other sciences, when the results of repeated measurement differ among themselves, it is customary, in the hope of eliminating or reducing their errors, to take a weighted average, and to employ linear equations of the following type :—

$$X_{gj} = W_1 X_{1j} + W_2 X_{2j} + \ldots + W_n X_{nj} \equiv \sum_1^n (W_k X_{kj}) \ldots \text{(i)}$$

Here X_{gj} will represent the best obtainable estimate for the true mark of candidate j, X_{kj} the marks actually awarded to him in turn by each examiner k, and W_k the best multipliers for weighting each examiner's marks ($k = 1, 2, \ldots n$). If there are N candidates there will be N equations of this type ($j = 1, 2, \ldots, N$).[1]

To determine the weights, the method of least squares is commonly adopted ; and the solution of the n linear equations resulting leads, as is well known, to a formula that enables each weight to be written as the ratio of two determinants. In the present instance the solution[2] will be

$$W_k = (-1)^{k+1} \frac{\Delta_{gk}}{\Delta_{gg}} \qquad \ldots \text{(ii)}$$

The determinants are based on a system or matrix of numbers that is variously written

$$\begin{vmatrix} \Sigma(X_g^2) & \Sigma(X_g X_1) & \Sigma(X_g X_2) & \ldots & \Sigma(X_g X_n) \\ \Sigma(X_1 X_g) & \Sigma(X_1^2) & \Sigma(X_1 X_2) & \ldots & \Sigma(X_1 X_n) \\ \Sigma(X_2 X_g) & \Sigma(X_2 X_1) & \Sigma(X_2^2) & \ldots & \Sigma(X_2 X_n) \\ \ldots & \ldots & \ldots & \ldots & \\ \Sigma(X_n X_g) & \Sigma(X_n X_1) & \Sigma(X_n X_2) & \ldots & \Sigma(X_n^2) \end{vmatrix} \text{ or } \begin{vmatrix} r_{gg} & r_{g1} & r_{g2} & \ldots & r_{gn} \\ r_{1g} & r_{11} & r_{12} & \ldots & r_{1n} \\ r_{2g} & r_{21} & r_{22} & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{ng} & r_{n1} & r_{n2} & \ldots & r_{nn} \end{vmatrix}$$

$\equiv \mathsf{M}_S$ (say) $\equiv \mathsf{M}_R$ (say) ... (iii)

—the summation being now from 1 to N.

Δ_{gk} and Δ_{gg} denote, as usual, the minors of the elements r_{gk} and r_{gg} respectively in the determinant M_R, or of the corresponding

[1] In adopting double suffix notation I shall put the suffix indicating the examiner or test first (e.g., k) and the suffix indicating the candidate or person tested (e.g., j) last.

[2] cf. Whittaker and Robinson, *Calculus of Observations* (1924), pp. 231 and 342. The elements in the determinants are usually given in the form of product-sums in earlier works on the method of least squares ; and in the form of correlation coefficients in proofs of the partial regression equations with which psychologists are more familiar : e.g., Kelley's *Statistical Method* (1923), p. 296. I myself should prefer to start with mean product-sums, i.e., covariances. (The terms in the two forms of equation (iii) are more fully defined in para. 567.)

elements in M_g. In the present inquiry, however, we are confronted with a somewhat exceptional problem : generally, with psychological tests, as in physical measurements, there is, for the simplest cases, an independent criterion for g; here there is none.

To the practical worker, errors are merely quantities of small theoretical importance which have to be recognized only in order to eliminate them. To the psychologist, errors, particularly human errors, have themselves become a central object of study. Accordingly, if we are to understand the causes of inaccuracy in examinations, we shall be forced to concentrate our attention on the system of relations symbolized by the set of numbers from which the determinants just indicated will be derived in any given instance.

In other sciences the problem would at once be visualized in geometrical form. We have a number of variables measured along axes that deviate more or less in their general direction. We choose as our main co-ordinate an axis that best represents the general direction of them all, and resolve each measurement into terms of its projection upon this central or centroidal axis. To take a concrete analogy, the examiner can be pictured as an aeroplane or a planet moving across the sky ; by his perception of the true values he is driven mainly in a given direction, but his course is more or less displaced by the numerous irrelevant influences to which he is subject. Thus the whole problem before us is analogous to that of the resolution and composition of vector quantities, such as forces in space, or the " generalized forces " of chemistry and thermodynamics.[1]

549. To those unfamiliar with these more technical devices the problem may be rendered clearer by turning to the inverse line of approach. Instead of trying to derive a hypothetical true mark from the actual mark, let us see how the actual mark is built up from the hypothetical.

Let us begin with the simple assumption that the total mark awarded by each examiner for each script can be treated as the

[1] Forces producing motion in space seldom involve more than three dimensions or co-ordinates. But to describe the dissociation of one gas into another—nitrogen tetroxide into nitrogen dioxide, for example—we may require to know the volume, the pressure, the temperature, and the mass of each gas present : for some purposes each molecule of gas may require to be treated as an independent variable, so that, if there are N molecules, we may have at least $5N$ variables to consider. Their relations therefore may be described by a system of $5N$ co-ordinates in generalized space : (cf. Jeans, *The Dynamic Theory of Gases*, 1922, pp. 35 *et seq.*) Thus what has come to be called " statistical mechanics " probably provides the nearest physical analogue to the kind of multiple factor analysis that is under consideration here. (See also note on matrix methods in modern physics, p. 312 below.)

sum of marks given for separate elements, each element being variously weighted by the various examiners according to the differences in their criteria or their judgment. Such an assumption is most natural in the common type of examination-paper in which a prescribed number of questions are to be answered, and each question calls for a definite number of statements: (e.g., " Answer 10 questions," Question 8 being " Give the third person singular, imperfect subjunctive, of " ten French verbs). A particular candidate j may give 6 statements out of 10 correctly; and a particular examiner k may allot 5 marks out of 100 to that question; then j's mark from k for that question will be $5 \times \frac{6}{10}$.

Here are two varying quantities: (i) the maximum allocated by the individual examiner to each question, which acts as a proportionate weight (r, let us say); and (ii) the proportion of that maximum gained by the individual candidate (z, say). If there are q questions, there will be q such products to sum. Accordingly, we may put the total mark awarded by k to j

$$X_{kj} = r_{k1} z_{1j} + r_{k2} z_{2j} + \ldots + r_{kq} z_{qj} = \sum_{1}^{q} (r_{ki} z_{ij}) \quad \ldots \text{(iv)}$$
$$(i = 1, 2, \ldots, q).$$

As before, if there are N candidates there will be N such equations for k; and, if there are n examiners, there will be n such equations for each candidate.

In other examinations each question or even the whole paper may call, not for a string of separable statements, but for one brief essay, forming an organized whole; and the marks will be probably awarded more or less by a general impression. Yet even here we may suppose that the total impression is the resultant of separate impressions in regard to distinguishable elements—qualities of style or logic, it may be, as well as pieces of information: and, as before, different examiners may implicitly weight the separate elements very differently, one, it may be, according no weight whatever to an element on which others place the greatest stress. In such a case, it will be more natural to treat z_{ij} as the ideal mark that an infallible examiner would assign to certain candidates for the ith element of work and r_{ki} as the proportion of the true mark which the fallible examiner tends to allot for such an element.[1] Accordingly, here we should rather analyse the mark given to j's answer to Question 8

[1] Or, more accurately, the proportion he allows on the average to all candidates who should have the same mark z, presuming both fallible and infallible examiner's marks to be reduced to the same scale.

TABLE 134

A Mark-Sheet as the Product of Two Matrices.

2nd Matrix (M_2)
Candidates' Hypothetical Marks for the Different Elements

	Cand. 1	Cand. 2	...	Cand. N
1st elt.	z_{11}	z_{12}	...	z_{1N}
2nd elt.	z_{21}	z_{22}	...	z_{2N}
...	...	...	...	...
qth elt.	z_{q1}	z_{q2}	...	z_{qN}

1st Matrix (M_1)
Examiners' Hypothetical Weightings for the Different Elements

	1st elt.	2nd elt.	...	qth elt.
Exr. 1	r_{11}	r_{12}	...	r_{1q}
Exr. 2	r_{21}	r_{22}	...	r_{2q}
...	...	...	...	...
Exr. n	r_{n1}	r_{n2}	...	r_{nq}

3rd Matrix (M_3)
Candidate's Total Mark for All Elements with Each Examiner

$\Sigma(r_{1i}z_{i1})$	$\Sigma(r_{1i}z_{i2})$	...	$\Sigma(r_{1i}z_{iN})$
$\Sigma(r_{2i}z_{i1})$	$\Sigma(r_{2i}z_{i2})$	...	$\Sigma(r_{2i}z_{iN})$
...	...	...	...
$\Sigma(r_{ni}z_{i1})$	$\Sigma(r_{ni}z_{i2})$	...	$\Sigma(r_{ni}z_{iN})$

NOTE.—In M_3 the "total mark" entered as the 1st term of the 1st row (= the top term of the 1st column) is the "inner product" of the 1st row of M_1 and the 1st column of M_2: and similarly for the other terms.

as $\frac{5}{10} \times 6$. But with either type of marking each constituent mark may be regarded as the product of two quantities—a coefficient and a variable.

550. Now these elements need have no real existence. They exist primarily for purposes of description and measurement, like North, South, East, and West, or the lines of latitude and longitude. A psychologist analysing the candidates' abilities would naturally seek for components that are in some sense real, that is, identifiable in the concrete. But the mathematician, making a purely quantitative analysis, will realize that he is merely reducing one set of variables to another set that seems to him simpler or more convenient to work with.[1]

551. The products are summed for the whole script; and on these principles we may in theory represent any mark-sheet, such as those investigated in this volume, as a set of numbers (M_3 say) which is itself the result of two other sets—the examiners' weighting coefficients (M_1) and the candidates' varying marks for the hypothetical elements (M_2): (*see* Table 134). The relations between such sets can be most easily handled by means of matrix notation[2]; and the rules for compounding two of them follow,

[1] What is tentatively assumed is not so much the existence of the elements, but the postulate that, to a first degree of approximation, they may be treated as additive: i.e., that each is itself compounded by the summation of units (otherwise they could not be measured), and that, when they come together, the resultant is a weighted or unweighted sum. Could Bacon's *Essays* be compared with Macaulay's on such a basis? Are not mental elements simply aspects of something that is alive and growing, fragments that are integrated organically into a pattern which can never be treated as the sum of the separate parts? To my mind such questions form a warning rather than an objection: by the choice of some more complex mathematical function and the use of imaginary quantities, we could in theory, I presume, deal with the problems of psychology as accurately as with those of any other concrete science. Addition we may regard merely as the simplest means of obtaining a first approximation.

It is quite conceivable that for certain purposes the relation of two elements might be better represented by some other function, e.g., by multiplication. Thus, if one candidate's memory is twice as good as a second's, this might lead the first to introduce twice as many facts into his answers in an intelligence test, so multiplying his mark rather than simply adding to it. Similarly, the fact that an examiner discriminates qualities in the various candidates that the other examiners have missed tends to multiply his marks rather than add to them. With the analysis suggested below, however, this could easily be allowed for: the discriminative peculiarity of the examiner, which at first sight looks like a "unique" or "singular" factor causing additions, would be treated as a multiplying factor, altering the unit of his scale.

[2] The suggestion that such problems as the present may be most conveniently treated by the use of determinants and matrices I owe primarily to Dr. W. F. Sheppard's brief chapter on statistics in his little book on *Determinant and Tensor* (1923, pp. 92–114: cf. also Turnbull and Aitken, *The Theory of Canonical Matrices*, 1932, pp. 173–6). Here I have not ventured to adopt tensor notation: but I may note that the influence of errors, which may be regarded as producing the effects of strain in the frequency-surfaces, can be dealt with most effectively by that device. As matrix algebra is unfamiliar to many students, I have thought it wiser to add short alternative proofs for the simpler cases along the usual lines of correlational statistics.

it will be observed, the usual rules for multiplying determinants and matrices: in short, $M_3 = M_1 \times M_2$. Briefly, then, our essential problem is this: given a set of total marks (3rd matrix), to discover the set of weightings (1st matrix) which will best account for it.

The method of reconstruction, like the method of analysis, thus leads once again to relations that can be represented geometrically. The ordinary Cartesian co-ordinates $\{x, y, z\}$, for example, form a matrix of one column (or vector), denoting a point in a space of three dimensions; and, if there were only three elements in a script, we could represent each candidate by such a point, and his mark by the corresponding vector. Similarly, more extended matrices such as those in Table 134 may be treated as representing co-ordinates in a space of many dimensions.[1]

552. This is in conformity with the general principles implicit in all attempts at mental measurement by means of multiple tests. For purposes of such measurement, as I have suggested elsewhere,[2] the mind may be regarded as a multi-dimensional continuum. Each mark or test-measurement will then represent a quantity measured in a definite direction. These directions may be infinite in number. Hence it will be of great practical convenience if we can resolve them into a smaller number of independent dimensions or axes, chosen to serve as an orthogonal frame of reference. If we like, these dimensions can be regarded as symbolizing fundamental aspects or capacities of the mind, fundamental merely in the sense that they afford the greatest amount of simplification.

Accordingly, we assume that any given candidate or testee is characterized by a set of hypothetical measurements measured along each of these fundamental dimensions. Diagrammatically he will be figured, in accordance with the usual convention, not as a multi-dimensional solid, but as a point with assignable co-ordinates in multi-dimensional space—the corner of his parallelepiped. For every candidate each test will measure, in the form of a vector-quantity, a performance which is a

[1] The first to suggest the systematic application of the methods of analytical geometry to the analysis of mental factors, and to formulate what he termed the cosine law and the vector equations, was apparently Maxwell Garnett (*Proc. Roy. Soc.* (A), XCVI, 1919, pp. 102–5). The principle, however, is implicit in Bravais's well-known deduction of his famous formula for correlation which was originally demonstrated as a cosine law ("Sur les Probabilités des Erreurs de Situation d'un Point." *Mémoires de l'Institut de France*, IX, 1846, pp. 260 *et seq.*).

[2] "The Mental Differences between Individuals," *British Association Annual Report*, 1923, Section J (Presidential Address, pp. 123–6).

complex resultant compounded of the fundamental capacities; and each test will weight the several capacities in different degrees. The outcome will be the test-measurements or marks as actually observed; and these will enable us to give to each testee his individual position in the imaginary space.

The object of psychological analysis, therefore, will be to resolve these actual measurements into their theoretical components, obtaining first the weightings implied by the several tests and thence the hypothetical measurements for the individual testees.[1] The problem is evidently one to be attacked by the methods worked out for the transformation of co-ordinates in analytical geometry.

553. The two modes of inquiry which we have been following, however, here lead to slightly different diagrammatic constructions. The two geometrical representations which they suggest may be briefly outlined at the start, because, as we shall see, the difference between them is accountable for the apparent divergences between the two main methods of factor-analysis hitherto proposed. The difference turns on the question: which shall we take for our first set of axes—the several examiners (or tests) with their actual marks, or the various hypothetical components? Let us call the two alternative constructions that of T-axes (test-axes) and F-axes (factor-axes) respectively. Of the two, the former is more in keeping with the conceptions

[1] Or we might take tests as points and persons tested as vectors, thus inverting the characteristics of M_1 and M_2. (As it is, we assume that the rows of the former, but only the columns of the latter, are uncorrelated, and the sums of the squares in each row of the latter add up to unity.) The importance of this further investigation will be realized if we think, not of examiners marking the same scripts in the same subject, but of examiners (or tests) for different subjects. In such a case, we should first translate the marks into terms, not of the standard deviation for the same candidate (as is done below), but of the standard deviation for the same examiner: then, instead of correlating the marks given by pairs of examiners, we should correlate the marks obtained by pairs of candidates. This would lead to a grouping of the candidates according as they resembled or diverged from the general type. With the subjects of the elementary school something of this sort has already been attempted. Pupils, for example, have been correlated according to their relative abilities in, and their relative preferences for, particular branches of the curriculum. Thus, when correlating scholastic subjects or tests, we are classifying the abstract abilities according to their common components; when correlating the pupils tested, we are classifying the concrete individuals. And generally the correlation of tests leads to an analysis of the human mind in the abstract; the correlation of testees leads to an analysis of the concrete human population. Or, to borrow the terms proposed by Stern (the first to distinguish clearly the two lines of approach), the first method uses "covariability" to study "intervariability" and the other uses it to study "intravariability." In the final analysis, through the old M_2 or the new, we might hope to reach a description of each man's genetic constitution. (By way of illustration I may refer to my Memorandum for the Board of Education's Report on *The Primary School*, 1931, App. iii, pp. 277 *et seq.* Cf. Stern, *Differentielle Psychologie*, 1911, p. 259.)

of the statistician; the latter may seem more natural to the mathematician who has applied analytical methods to other fields.

Consider first the simplest possible case—that of only two examiners or tests whose agreement is attributable to a single component which both of them share. I shall assume that the examiners' scales have been rendered comparable by taking the average of each as origin, and his average variability as unit.[1] If we follow the first line of approach (that of trying to resolve rather than reconstruct our data; cf. para. 548, p. 247 above) we should adopt the plan employed in most statistical textbooks to illustrate what is called the correlation between two variables[2]; and we should plot the marks for each candidate on a two-dimensional diagram, measuring the marks for one examiner along the horizontal axis and those for the other examiner along the vertical. Each candidate will be represented by a single point; and equal frequencies will be represented by a series of similar and concentric ellipses. The longest diameter[3] of the ellipses will run obliquely between the two rectangular test-axes; and, since the units of the examiners' scales (their "standard deviations") have been equalized, the diameter must always bisect the right angle between those two axes, and so make an angle of $\frac{\pi}{4}$ with each. The extent to which the two examiners agree will be shown by a reduction in the transverse scattering of the points on either side of this line, i.e., by an increase in their density. When the two examiners are in absolute agreement, every candidate will be crowded on to this oblique diameter. Accordingly this may be taken as the axis along which we may measure the common component responsible for the resemblance between the two examiners; and the amount of variability along this line will be indicated by its relative length. Thus, if V_1 and V_2 be the lengths of the two semi-diameters of one of the ellipses, θ the eccentric angle of the point at which the

[1] See below, Section III, pp. 267–9.

[2] e.g., Yule, *Introduction to Statistics* (1930), p. 320. Brown and Thomson, *Mental Measurement* (1925), p. 122. cf. also Whittaker and Robinson, *loc. cit.*, p. 321 (multiple correlation derived from the expression for one set of variables given as linear functions of a second set). It is generally assumed that the frequencies are distributed approximately in accordance with the normal probability integral and that the regression-lines are linear: but this is not indispensable. Adopting matrix methods most of the essential results can be deduced without reference to the theory of probability.

[3] I use the phrases "longest" and "shortest diameters" as equivalent to "major" and "minor axes" to avoid employing the same term "axis" in a confusing number of different senses.

ellipse cuts the vertical axis, and $r_{kk'}$ the resemblance between the examiners, it is easy to show

$$\frac{V_1}{V_2} = \frac{tan\ \theta}{tan\ \frac{\pi}{4}} = \frac{\sqrt{1+r_{kk'}}}{\sqrt{1-r_{kk'}}} \text{ or } r = \frac{V_1^2 - V_2^2}{V_1^2 + V_2^2} = -cos\ 2\theta.$$

The shortest diameter will accordingly indicate the relative amount of error. But such a diagram, it will be observed, does not distinguish between the separate tendencies to error that are affecting the two examiners, i.e., between the two distinct and individual components which are reducing their correlation from unity.[1]

554. If we follow the second mode of approach (F-axes), we shall take as our vertical and horizontal lines of reference, not the lines for the examiners or tests, but the lines for the hypothetical components. Since the errors of the two examiners are necessarily independent (everything that is shared having been embodied in the common factor), we shall need, not two rectangular axes, but three ; and what appeared as an ellipse on a two-dimensional plane will be inflated into a three-dimensional ellipsoid.

If, further, we assume that the units for these three hypothetical components are equal, i.e., that the average variability is the same in all three dimensions, the ellipsoid will be strained and stretched along its minor axis until it becomes a sphere. The lines representing the two examiners will now appear as two axes, more or less oblique, passing through the centre of this globe like knitting needles through a ball of wool, but still in vertical planes at right angles. When the examiners agree completely with the common component, their lines will coincide with the upright central axis. As they depart from it, their lines will incline further away from the vertical. In short, the direction cosines of their respective lines with reference to this vertical axis will measure their agreement (r_{kg} and $r_{k'g}$) with the common component which it represents—i.e., they will indicate the relative weight that each examiner is unconsciously giving to the hypothetical true mark. And similarly the agreement between the two examiners themselves ($r_{kk'}$) will be measured by the cosine of the angle between their lines. If both are pulling in

[1] If we are dealing with an examination in two different subjects rather than with two different examiners, the second component in either case may be regarded as consisting chiefly of an irrelevant ability, measured, like all the components, by a plus quantity if above the general average, and by a minus quantity if below, and peculiar to that subject: thus, if we are trying to assess children's intellectual capacity by a scholarship examination in arithmetic and English, the relative failure of one child in arithmetic alone may be regarded as due to some peculiar incapacity of his, limited to that subject and acting as a " negative error "—i.e., as an irrelevant tendency dragging down his total mark towards or even below the general average.

the same direction, the amount of agreement will be + 1·00, i.e., positive and complete; if each is pulling at right angles to the other, it will be zero, i.e., they will neither agree nor disagree; if each is pulling along the same line, but in opposite directions, it will be — 1·00, i.e., the disagreement will be so complete that one examiner is in effect simply reversing the order of the other. Moreover, the cosine of the angle between the two test-lines will be the product of the direction-cosines of the two test-lines themselves: here, therefore, since the test-lines lie in planes at right angles (being so far assumed to be dependent on two components only) $r_{kk'} = r_{kg}\, r_{k'g}$.

Thus with T-axes the oblique axis is fixed, and the extent of the agreement is shown by the varying shape of the ellipses; with F-axes the shape of the ellipses is fixed (being in fact circular), while the extent of agreement is shown by the varying angles of the oblique axes.

555. The results deduced from the two alternative constructions can readily be reconciled when we note the relation between the two. The construction drawn with T-axes is simply a projection on to a plane of n dimensions of the construction made with F-axes which is in $(n + 1)$ dimensions; and here $n = 2$. With F-axes the variability of all three components is made equal, each to each, and therefore equal to that of the two tests. The contours for the frequencies thus become circles lying in the oblique plane that passes through the two test-lines in the three-dimensional diagram. The ellipses in the two-dimensional diagram with T-axes are thus the shadows, as it were, of these circles projected on to the plane which contains the test-lines there. In other words, the T-diagram is the F-diagram seen from above and re-drawn on the flat.

In the three-dimensional construction the plane of the test-lines is tilted above the horizontal plane at an angle whose cosine is

$$\frac{\sqrt{1 - r_{kg}\, r_{k'g}}}{\sqrt{1 + r_{kg}\, r_{k'g}}} \quad \text{that is} \quad \frac{\sqrt{1 - r_{kk'}}}{\sqrt{1 + r_{kk'}}}$$

Hence, in the two-dimensional diagram, the minor axis of each ellipse will show a foreshortening in that proportion; and in the same diagram the unit of variability for the tests will be a line which is simply half the projection of the ellipse itself on either of the test-lines. It follows that on the T-diagram the correlation could be measured, as we have just demonstrated, by the cosine

of the complement of double the eccentric angle of the ellipse at the point where it cuts the vertical axis. The expression more frequently taken is the tangent of the angle between the regression-line (the line bisecting the horizontal chords of the ellipse) and the vertical axis. The equivalence of these two formulæ for the same correlation sums up the relations between the two constructions.[1]

556. Now consider what happens when there are more than two examiners or tests. The two simple geometrical figures will become intricate constructions in multi-dimensional space. With T-axes there will be as many dimensions as there are examiners or tests ; with F-axes as many as there are components or factors. In the one case the distribution will assume the form of an n-dimensional football, and in the other that of an $(n + 1)$-dimensional cricket ball. And in this hyperspace each candidate will still be represented by a point. With the T-axes the rectangular co-ordinates of his point will again represent his actual marks. With the F-axes they will represent his hypothetical marks for each component ; and his actual marks as received from the examiners will be represented by the projection of his vector-distance on each examiner's line in turn. But the ensuing and outstanding difference will remain as before : we shall find the influence of the common factor producing in the one case ellipsoids of varying form and density, and in the other oblique axes variously inclined. And again the results of the two alternative constructions can be reconciled by treating the first as derived from the second by a process of stereographic projection or the second as derived from the first by a process of homogeneous strain. The formulæ for the necessary transformations are easily obtained from n-dimensional geometry or the algebra of quadratic forms.[2]

[1] Expressed in words the explanation may sound a little complicated ; it will, however, become quite clear if rough geometrical constructions on paper are attempted for the simpler cases.

[2] The former construction, as we shall see, is the geometrical picture suggested by Hotelling's analysis (" Analysis of a Complex of Statistical Variables," *Journ. Educ. Psych.*, XXIV, 1913, pp. 422 *et seq.*, pp. 417–441, 498–520). The latter is implied by Spearman's (*Abilities of Man*, 1927, Appendix, pp. i–xxiii ; it should be observed that Spearman's highly original work has formed the starting point of almost all the mathematical investigations upon this and kindred problems). Thurstone (" Multiple Factor Analysis," *Psych. Rev.*, XXXVIII, 1931, pp. 406–427) extends it still further by first taking as his co-ordinates, not one line common to all planes and n other lines independent of the first, but any number m of independent lines which may be shared by several planes, leaving the axis common to all still to be discovered. Although at first sight the different methods seem to lead to divergent results, the final formulæ can in point of fact be easily related one to another on the principles indicated above.

557. Of these two modes of representation I shall here, in the main, adopt the second, as leading to somewhat simpler solutions.

We shall accordingly begin by postulating that the ultimate components we are in search of shall be (1) mutually independent and (2) as few in number as is practicable, i.e., that our marks shall be reduced to terms of rectangular co-ordinates in the simplest possible space.

The first requirement is easily fulfilled. It is well known that points defined with reference to n oblique axes can be described as points in a space defined by the same number of orthogonal axes : the transformation is effected by means of a matrix showing the direction cosines of the oblique axes as referred to the orthogonal. Applying this principle to the results of mental testing, Garnett pointed out and formally proved that any n correlated measurements, such as are provided by n tests, can always be expressed in terms of n independent hypothetical components. But these may be selected with $\frac{1}{2} n (n - 1)$ degrees of freedom.[1] When we have decided which of these many frames of reference is the one most appropriate for our purpose, a simple rotation of the axes will effect any further transformation that may be needed.

The second requirement is facilitated by a tendency that is empirically found to hold good in many, if not most, mental measurements of the kind we have to consider. When a matrix of weightings for a set of hypothetical components has been more or less accurately determined, it generally appears that, for one and possibly more of the components, the weighting, though large, is of much the same order with every examiner or test : it varies, but not very widely ; whereas for each of the remaining components the weights are exceedingly small with all but a limited group of examiners or tests.

558. Accordingly we may reduce the component elements to the four following categories : (i) those which *every* examiner is treating as relevant ; (ii) those which only *some* of the examiners are treating as relevant ; (iii) those which only *one* examiner is treating as relevant ; and to these we must add (iv) any elements which may arise from the many minor accidental influences affecting the examiner unconsciously, and which are therefore unrelated to elements actually present in the candidate's work : in a word, from the effects of chance.

It will be convenient to have a name for each of these components. I shall speak of them as " factors," borrowing the

[1] " The General Factor in Mental Measurements," *Brit. Journ. Psych.*, X, 1920, pp. 243–4.

term from current psychology,[1] and shall classify and designate them as follows[2] :—

- A. *Common* (factors influencing several examiners or tests).
  - 1. *Universal* factors (if one, usually called "*the* general factor") : *G*.
  - 2. *Particular* factors (variously called specific, overlapping specific, group factors, or general factors of limited range) : *S*.
- B. *Individual* (factors peculiar to one examiner or test).
  - 3. *Singular* factors (variously called specific, unique, or individual factors; constant or systematic errors) : *U*.
  - 4. *Chance* factors (random errors or unreliability) : *E*.

Consequently, we may now re-group the terms on the right-hand side of our initial equation (iv), and express the mark of any given candidate as follows :—

$$X = \Sigma(r_{kg}.\ G) + \Sigma(r_{ks}.\ S) + \Sigma(r_{ku}.\ U) + \Sigma(r_{ke}.\ E) \quad \ldots \text{(v)}$$

[1] The mathematical reader will consider it unfortunate that statistical psychologists use the term "factors" for the hypothetical constituents into which they analyse a mental achievement, and then proceed to treat the "factors" exclusively by the method of addition. "Element"—the term proposed by Sir Philip Hartog—seems the word most appropriate to the constituents reached by a psychological or non-statistical analysis—e.g., such items as knowledge of this fact or that method, ability to apply this mode of logical argument or to adopt that form of literary expression, and the term "component" to the purely hypothetical variables reached by mathematical resolution. However, phrases like "general factors," "group factors," "specific factors," "factor-analysis," are now so well-established that it is probably too late to amend the terminology. Moreover, the number that is essentially characteristic of the "factor" is, as we shall see, treated as a coefficient.

[2] The terms that I have proposed are borrowed from traditional logic. "Universal," "particular," and "singular" propositions are those having the form "all," "some," or "this" "examiner(s) recognize such and such a quality," respectively. "General" and "specific"—the more usual terms in psychological writings—are relative terms; and therefore frequently ambiguous. A general factor means sometimes a factor common to all tests that can possibly be conceived, sometimes a factor common to a particular group. A specific factor means sometimes a factor common to a limited number of tests, sometimes a factor peculiar to one only.

It will be observed that a "universal" factor and a "singular" factor really form extreme cases of the "particular" factors. Thus, a universal factor, shared by every member of a Board of ten, might prove to be only a factor of limited generality, if we added an eleventh examiner who did not happen to possess that factor. Similarly, a "singular" factor, peculiar to one examiner only among the Board of ten, might be shared by an eleventh and so turn out to be "particular." Hence, if "universality" and "singularity" are defined by reference to one set of tests or one set of examiners only, they must remain relative to the composition of that set. So far as intellectual tests are concerned, it usually turns out that, if the sets are suitably chosen, a factor universal in one is generally universal in another; and the same seems to hold good of examiners. On the other hand, by choosing tests which are sufficiently similar—for example, if one test is almost a repetition of another—a singular factor can always be converted into a particular. But, as a rule, its range remains remarkably narrow.

559. The same analysis would serve for almost any form of measurement. Thus, suppose our examiners were school medical officers recording measurements of the examinees' weight; we could in theory subdivide each crude figure into the sum of four analogous components: (i) the figure for the total weight of the child's naked body—the quantity which will mainly determine every doctor's measurement—a "universal" factor; (ii) an additional figure for the coats, or shoes, or underwear which some doctors allow to be worn and others do not—a "particular" factor; (iii) a constant error due to defects in an individual doctor's weighing-machine or to his special method of using it —a "singular" factor; (iv) a random error due to coarse or careless readings, erratic variations in each child's state of health, meals, exercise, and the like—a "chance" factor.

When we turn from physical measurements to those of mental capacities or attainments, a further difficulty confronts us. In comparing physical magnitudes, we assume that all observers are using the same scale: if a French doctor takes a temperature with a Centigrade thermometer, while his English colleague records the result in Fahrenheit, we make the necessary conversion before the two readings are compared. Similarly, in dealing with marks obtained in an examination, we must also take into account possible differences of scale—that is, differences of implied unit and differences of implicit zero-point or mean.

560. Thus I suggest that an examiner's actual marks may be regarded as the result of six main characteristics:—

(i) *The standard of severity.* In mental measurement, as in spatial measurement, there are no minimal points which can be taken to mark an absolute zero. The sea-level is not the bottom of the land, but a convenient average. In the same way, the best criterion for each examiner's general standard of marking will be, not an imaginary nought on an exclusively positive scale, but the average mark which he is awarding to a large and typical batch of candidates. This we shall treat as the origin or arbitrary zero from which to measure positive and negative variations. An examiner whose average for a whole series of scripts is lower than another's may be described as more severe; this is a characteristic that can be readily determined, and if necessary discounted, by first computing for every examiner the average or arithmetic mean of all the marks he allots.

(ii) The *distribution* of the marks about the average. This really involves two characteristics, of which only one will be taken into account here.

(a) The general range, that is, the extent to which the marks

are spread out above and below the average. The difference is essentially a difference of unit. One examiner may mark on a scale of 0 to 100 and another on a scale of 0 to 200. Or both may mark on a scale with the same nominal maximum and even the same actual average, 50 let us say; yet one may mark his top and bottom candidates 60 and 40 respectively and the other 80 and 20. The effective limits of the latter's scale are thus twice as wide.

This source of discrepancy is quite as serious as a difference in the general average; yet it is far more easily overlooked.[1] The *extreme* range, however, is an unsatisfactory criterion. The *average* range, that is the mean variation about the average, is better, since it takes into account all the marks awarded. The best and most usual is the root mean square variation, commonly termed the *standard deviation*.

In what follows, the standard deviation, denoted by σ, will be adopted throughout as the general measure of variability, and will be used as a comparable unit. The square of the standard deviation—the mean of the squares of the individual variations—is termed the variance.[2] Variance is additive. Thus, if g, s, and e represent marks for three independent qualities combining to produce a single composite result, x,

$$\sigma_x = \sqrt{\sigma_g^2 + \sigma_s^2 + \sigma_e^2},$$

as in resolving a polar co-ordinate into a number of rectangular co-ordinates. If the standard deviation (σ_x) is put equal to 1, then, with n examiners, the total variance will obviously be n.

(b) The *curve of distribution*. Of several examiners, all keeping to the same average and all showing the same standard deviation, one may be very generous with his high marks; another with his low marks; a third may award nearly an equal number of

[1] In particular, it should be noted that, unless the marks of the several examiners are in perfect correlation, the spread or standard deviation of the marks derived by averaging those awarded by several examiners is nearly always less than that of the marks of each examiner taken separately: thus, if each examiner means to allot one distinction and one failure in the marks for his own subject, it will often happen that (so long as the same borderline is preserved) no one gets a distinction or a failure when the marks for all subjects have been added or averaged.

This is a point constantly ignored by examining boards which mark numerically instead of by letters: we shall meet the same difficulty later on in our analysis when we sum the factors influencing each particular examiner. The effect can be quite easily calculated and allowed for.

[2] If each individual is conceived as a point in space, the "frequency" or number of individuals may be conceived as a density: hence, by a natural analogy from dynamics, the variance is sometimes described as the "second moment" about the mean—the deviations being regarded as distances and the frequencies as weights attached or concentrated at those distances. The parallel may be suggestive to the physicist, and will be used again in a later argument. On the difference between Dr. Rhodes' treatment of the standard deviation and my own, see p. 312; also note 1, p. 272.

each, but very few medium marks ; and so on. Such differences will gravely affect the number of failures and of first classes or distinctions that the different examiners award. Here I shall assume that the distributions are approximately "normal".[1]

The enumeration of the remaining characteristics will be in its essence an attempt to analyse the variance. The total variability shown by the n examiners is to be divided into four contributory portions—a "common," a "partly common," an "individual," and a "residual" variance : these when summed together must make the total n.

(iii) The *true value* of the candidates' work—for our purposes the most important component of all. No external or objective criterion is available : we can only extract an estimate from figures on the actual mark-sheet. Hence the true value is a purely abstract and hypothetical concept ; but so are all the concepts of quantitative science—like the atomic weight of an element or the direction of the true North. How is it to be defined ?

The simplest definition would consist in identifying the true list of marks with the universal or general factor, that is, with the effect of the element or elements recognized by all examiners. Though it has no objective existence, this ideal mark-list must operate as a common and all-pervading agency influencing each of the examiners, but influencing them in different degrees. To this definition, however, it may be objected that the elements recognized by all the examiners might be comparatively few in number. Ought we not, therefore, to incorporate any and every element that may be recognized by a majority, even though it is a different majority for each of the different elements ? In answer it must first be recalled that the subdivision into factors is not quite the same as the subdivision into elements : we have in effect slightly rotated our axes. As we have seen, the general elements are very general ; and the specific are very specific. If the examiners are numerous, those elements that are common to nearly all of them, but not quite all, will be treated as universal elements, having a very low weight with one or two. Accordingly, we may slightly amend the simpler definition, and describe the true mark-list as that which would show the least amount of general disagreement with the mark-lists of all the examiners. The phrase "least amount of general disagreement" may be

[1] Differences in the form of distribution can be brought out by plotting the marks on a graph to show the frequency-distribution in the shape of a curve. By taking higher powers of the deviation (higher "moments" in addition to the squares) they could be reduced to a mathematical form. But such constants have large probable errors ; and an examination of their nature would here be out of place.

interpreted in accordance with the principle of least squares; and this will be equivalent to saying that the true mark-list is that which has the highest mean square correlation with the actual mark-lists taken in turn. Such a definition will imply that we identify the set of true values with the hypothetical component that contributes most to the total variance.[1]

(iv) *Limited influences.* In most examinations the irrelevant factors that are likely to bias two or more examiners in the same direction will be fairly obvious. In essay papers, such things as spelling, grammar, handwriting, verbal expression, literary style, may count more with some examiners than with others. In subjects that involve questions of taste or doctrine rather than of fact or logical deduction—in art, literature, philosophy, for example, as distinct from languages, sciences, and mathematics—the particular school of thought towards which certain examiners lean, or the particular prejudices which two of them share, may make one examiner's marks agree unduly with a second's, and seem positively antagonistic to those awarded by a third. But these are not the only tendencies that are likely to bias a man's marking.

(v) *Personal influences.* There are other influences more elusive and less easy to detect, because they are peculiar to each single examiner. In the main these are likely to be a matter of personal feeling or emotion rather than of intellectual attitude or taste. Generally, it may be said that every influence inducing a given examiner to swerve from the true mark operates, like other irrelevant and irrational influences, more or less unconsciously. But the less unconscious influences—those that are "fore-conscious" to borrow a convenient term from Freud—are for the most part those which the examiner may share with other members of his group: they can, with a little effort and self-understanding, be consciously allowed for. The more personal influences are so deeply unconscious (in the psychoanalytic sense) that the plain man, no less than the psychoanalyst, realizes that it is always unsafe to trust to the judges' own powers of adjustment. As a surgeon is expected never to

[1] It might be objected that this will lead to a component which covers more of the total variance than the "general factor" as defined in the preceding paragraph (the definition commonly adopted by Spearman and his followers), and result in negative weightings for the residual factors. In point of fact, how much of the variance it covers will depend largely on the data included in our calculations: e.g., whether we include the socalled "reliability coefficients" in the initial correlational table. In any case, the application of Spearman's formulæ itself usually yields specific correlations with negative signs. This may be a drawback when we are dealing with mental abilities, for it is difficult to think of an ability as inhibiting performance instead of reinforcing: but a negative correlation between two examiners is by no means inconceivable.

operate on his own relatives or even to diagnose their more serious complaints, so, instead of accepting the estimates of a master or tutor who knows his pupils or his students at first-hand and is therefore bound to have his prejudices and his favourites, we call in an external examiner or appoint an external examining body. Much the same holds true of subjects : if an examiner has taken some special problem for his own private research or his personal writings, he will tend to be unduly interested and influenced by the extent to which a candidate reproduces his teaching, quotes his books, or prefers the view of an opponent.

(vi) *Accidental influences.* Finally, except in the most elementary of the abstract subjects—mechanical arithmetic, for example—there must, in every examiner's marking, be inevitably an ingredient of chance. By chance I understand the sum total of a very large number of very small influences, all irrelevant to the main purpose of the examination, and for the most part inseparable if not indefinable. Such miscellaneous influences as fatigue, lapse of attention, accidental changes of standard while working through a long series of scripts, will affect the marking quite irregularly if the order in which the papers are marked is unconnected with their merit (e.g., alphabetical order). A competent examiner will usually adopt some expedient for neutralizing these effects—for example, by going through the same scripts twice in a different order. But, even with the best precautions, the same examiner, unless he is guided by a retentive memory, will seldom give precisely the same mark on two successive occasions to precisely the same script. These fluctuations of the individual examiner about his own general estimate we may describe as his " random variation."

561. Thus the investigation of intellectual measurements becomes an attempt to discover factor patterns of the following kind (Table 135, where r denotes weighting coefficient as before and $\Sigma\,(r_{ki}^2)=1$).

TABLE 135

FACTOR PATTERN FOR THREE EXAMINERS

Examiner	Universal or General Factors			Particular or Specific (Group) Factors			Singular or Unique (Individual) Factors			Random Factors or Chance Errors		
	g_1	g_2	...	s_1	s_2	s_3	u_1	u_2	u_3	e_1	e_2	e_3
No. 1	r_{1g_1}	r_{1g_2}	...	r_{1s_1}	r_{1s_2}	0	r_{1u_1}	0	0	r_{1e_1}	0	0
No. 2	r_{2g_1}	r_{2g_2}	...	r_{2s_1}	0	r_{2s_3}	0	r_{2u_1}	0	0	r_{1e_2}	0
No. 3	r_{3g_1}	r_{3g_2}	...	0	r_{3s_2}	r_{3s_3}	0	0	r_{2u_3}	0	0	r_{1e_3}

Yet even now, it may be thought, the task will be unending. There seems no limit to the number of factors or elements in each of the four groups. Here, however, several considerations come to our aid.

First, the "universal" factors are distinguished amongst themselves only by relative differences in their weighting. Now it is a fact familiar to statisticians that, unless the number of variables to be weighted is exceedingly few and the correlation between the weights and the variables decidedly high, differences in weighting have but little effect: indeed, unless such differences can be determined with accuracy, their introduction impairs rather than improves the result. Hence, where merely approximate results are desirable or attainable, we may group or pool the various universal factors together, and so treat them as one: this single universal factor I shall refer to as "*the* general factor."

The same will hold good of the "particular" factors: if there is any one set that runs through the same particular group of examiners, the different weightings for each of them may be ignored. But we shall still have as many particular factors as there are combinations of examiners. Actual investigation, however, shows that, more often than not, the specific factors have not only a low weighting, but also an unexpectedly narrow range. Often they are shared by two examiners (or two examination subjects) only, rarely by more.[1]

Thirdly, for any given examiner and test it is statistically quite impossible to separate out more than one singular factor or more than one random factor. The influences peculiar to a single examiner combine into a single variance; and the same is true of the influences attributable to chance. Finally, unless the mark-sheet furnishes two sets of marks from the same examiner, his singular factor itself cannot be separated from his chance factor.

[1] This is conspicuously demonstrated in examinations in elementary school subjects: see Burt, *Distribution and Relations of Educational Abilities* (1917), pp. 52-62. The result is a kind of overlapping cyclic arrangement: it doubtless depends in part on the way we choose examiners or tests, or separate out the various subjects for the question papers set. If test B or subject B overlaps with A, we add another, C, which not only deals with a fresh field, but also trenches on that part of B not covered by A; similarly D will overlap with C; and so on, until finally Z overlaps with that part of A not covered by B.

The narrowness of range, and relative small weighting, found for all components after the first is also visible in analyses carried out by more recent methods. cf. Hotelling, *Journ. Educ. Psych.*, XXIV (1933), p. 434; Line, Rogers and Kaplain, *ibid.*, XXV (1934), pp. 58–65; Russell, *ibid*, XXVI (1935), p. 285; Thurstone, *Psych. Rev.*, XXXVIII, p. 498, *et seq.*; *id.*, *ibid*, XLI (1934), pp. 26, 29; Alexander, "Intelligence: Abstract and Concrete," *Brit. Journ. Psych. Mon. Supp.*, XIX (1935), p. 75; cf. also Kelley, *Crossroads in the Mind of Man* (1928).

Our problem is thus greatly simplified. Under the conditions imposed we can only subdivide the variance exhibited by a given examiner's marks into three main portions—a common, a partly common, and a residual variance.

562. Accordingly, we may now try to express the various influences we have enumerated by means of a single equation. Clearly we must add (or subtract) marks that represent the general standard, the general or universal factor, the specific factors (as we may call them), and the random variations, and multiply (or divide) for differences of unit. We may therefore write

$$\frac{X_{jk} - \overline{X}_k}{\sigma_k} = b_{kg} \cdot \frac{G_j - \overline{G}}{\sigma_G} + \Sigma(b_{ks} \cdot S'_j) + b_{ke} \cdot E'_{jk} \quad \ldots \quad \text{(vi)}$$

where most of the symbols have the same meaning as before; $\overline{X}_k$ is the average mark awarded by k to all the candidates; $\overline{G}$ their average true mark; the dashes affixed to S' and E' indicate that these terms, like the preceding, are to be expressed in standard measure; b_{kg}, b_{ks}, b_{ke}, may be regarded as new weighting coefficients, representing respectively the proportion of true, specific, and erroneous marks, included by k in his actual marks, when all are in standard measure.

Of all these constants the one of greatest importance is obviously b_{kg}. It sums up the reducing effect due to all the other influences, as though the examiner had to diminish the mark for true and relevant qualities to make room for the rest. We might, therefore, for the moment call b_{kg} his coefficient of accuracy. Were there no correspondence whatever between his actual mark and the true mark, then b_{kg} would be zero; but were he able, on the other hand, to eliminate the influence of S and E, and to conform exactly to the true average and the true dispersion, then b_{kg} would rise to 1. If all the examiners could succeed in doing this, their marks would be identical both with each other and with the ideal.

How far can we approximate to this perfection, or allow for lack of it in our intended analysis?

Section III.—Adjusting for Differences of Scale: the Average and the Standard Deviations

563. In practice it is possible to reduce the first three sources of error very considerably. The usual method is to attempt some definition both of the qualities to be marked and of the marks themselves, though such definitions are seldom based on a sound statistical foundation. One board of examiners, for example, tells its members that " marks should be sent in on a

five-letter scale ; ... *C* should represent a mark allotted to an average candidate : on a numerical scale, *C* is equal to 50 per cent." Such an instruction defines the average. Again, another examining body instructs its examiners that " 30 per cent. shall be the lower borderline for a bare pass, and 70 per cent. the lower borderline for those who merit distinction " : and adds that " in the past the average number of failures has been 4 per cent., and the average number of distinctions also almost exactly 4 per cent., of the total number of candidates presenting themselves at the examination." Such instructions implicitly define the amount of scattering—that is, the standard deviation—which the examiners' marks should show, and roughly indicate the general form of the distribution.

The best way to secure an effective agreement among the examiners in regard to all these three points is to insist that the marks should be distributed in broad conformity with a normal curve, and on that basis to define the letters or the numbers in terms of their expected frequency. A standard table of this sort fixes the average, the standard deviation, and the shape of the curve. This device has recently been introduced into certain examinations ; and, where a large body of examiners are concerned in the marking of a large number of candidates, has proved effective.

Even, however, where examiners do not conform to the definitions and injunctions of the examining body, or none has been laid down, it is still possible, by keeping the candidates in the same order but by readjusting the marks (e.g., adding marks to allow for too low an average, or multiplying the deviations to allow for too narrow a standard deviation) to correct in some measure the discrepancies arising from such sources.

564. The nature of these two outstanding influences, and the importance of allowing for them when they vary widely from one examiner to another, may be illustrated from the short mark-list already taken at the beginning of Part II to exemplify the methods of statistical analysis there adopted (see above, Table 117, p. 187). The averages and standard deviations[1] for each of the six examiners are shown in the following table.

[1] Calculated by the formula $\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{N}}$, where $x \equiv$ the individual's mark, $\bar{x} \equiv$ mean mark, and $N \equiv$ number in group. In dealing with estimates derived from small groups, and particularly when calculating probable errors or tests of significance, it is now usual to divide the squared deviations by $N - 1$ rather than by N. For the above table I have averaged them in the more ordinary way in order to keep the figures comparable with those given in the earlier part of this volume, where the " squares of random variations " have been averaged by simply dividing by N (para. 403, p. 190, and following tables).

TABLE 136

MARKS FOR SCHOOL CERTIFICATE, LATIN: GROUP I

AVERAGES AND STANDARD DEVIATIONS FOR THE SEVERAL EXAMINERS

Examiner	A	B	C	D	E	F	Average
Average Mark	38·60	45·07	49·80	39·66	40·33	42·47	42·46
Standard Deviation	3·74	3·89	3·87	5·30	4·16	3·81	3·83

It will be seen that the average of the most generous examiner, C, is more than 25 per cent. higher than that of the most severe examiner, A, and that the standard deviation of the most discriminative examiner, D, is more than 40 per cent. larger than the standard deviation of the least discriminative, A. As regards discrimination, all the examiners except D spread their marks to much the same extent; their range varies from 13 to 16 marks. D, however, stands apart; his range is 20 marks.[1]

The number of candidates in the list selected is small, namely, 15. But I have deliberately kept to a small sample in order to illustrate a point to which I personally attach considerable weight—the need for testing the statistical significance of the results obtained (whether by calculating the so-called "probable errors" or by applying some more appropriate criterion) before drawing any final conclusion as to methods or results. Here, with so few candidates in the group, the probable errors are inevitably high. Those for the two differences I have just cited are ± 1·12 (for the difference between A's and C's average) and ±0·79 (for the difference between A's and D's standard deviation).[2]

[1] It will be remembered that, in investigating the results of this particular examination, it was decided that the scripts of these 15 candidates should be "so selected that the candidates had obtained at the original examination exactly the same moderate mark" (see above p. 18). Unless, therefore, the correlation between the original examiners' marks and the present examiners' probable marks for the entire group is assumed to be zero, this procedure must have somewhat reduced the standard deviations, and their differences, below the figure that would have been obtained with a genuinely random sample.

[2] The formulæ used are those that are most familiar: $p.e._{diff} = \sqrt{p.e._1^2 + p.e._2^2}$; $p.e._M = \cdot 6745 \frac{\sigma}{\sqrt{N}}$; $p.e._\sigma = \cdot 707\, p.e._M$; a significant figure is one which is $3 \times p.e.$ Assuming a normal distribution, this gives, for the difference between A's and D's standard deviations, P (the proportion of cases in which the specified figure is likely to be exceeded by sheer chance) = ·14 or about 1 in 7. I should prefer to use standard errors rather than probable errors; but the latter are still almost exclusively used by psychologists.

With small groups, however, this procedure is apt to yield inexact results: for methods more appropriate to small groups see Fisher, *Statistical Methods for Research-Workers* (1934), pp. 120 *et seq.*, and 214 *et seq.* With any method of calculation the difference between the two averages remains significant: but the above procedure rather exaggerates what little significance might attach to the difference between the standard deviations. Applying Fisher's criterion I find P = nearly 1 in 5 instead of 1 in 7. (The criterion consists in taking $z = \log_e \frac{\sigma_2}{\sigma_1}$, and $s.d._z = \sqrt{\frac{1}{2}\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$, provided N_1 nearly = N_2 and is not too small; otherwise P for z_1, N_1, and N_2, can be read from the published tables, e.g., *loc. cit.*, pp. 232–5.)

The former difference is 10 times, the latter only twice, the size of its probable error. Thus, in any further analysis of this particular list, we should be forced to take into account the divergence between the examiners' general standards; but we might perhaps justifiably ignore, with a group so small, the differences between the standard deviations.[1] With a group of more than 35, however, the difference in D's standard deviation would begin to outweigh any difference between the averages: if a large batch of candidates were examined, D's exceptionally wide range would make far more difference to the number of failures or distinctions that he awarded than any of the observed divergences in general standard.

Accordingly, in analysing results from any larger mark-list, it would be eminently desirable to reduce each examiner's marks to terms of his own standard deviation before calculating the degree of his random variation. We can illustrate the effect from the mark-list before us; for C, for example, the average of the squares of random variations[2] would be reduced from 6·7 to 0·42, and for D from 6·9 to 0·21. With this correction, it will be observed, D, instead of being the examiner with the largest amount of "random variation," appears much less erratic than C. Actually, however, in view of the high probable errors inevitable in so short a list, it is to my mind scarcely legitimate to compare the individual examiners for Latin by means of any form of calculation: I cite the figures merely to illustrate the general procedure.

565. By calculating an examiner's actual average and standard deviation, then, we can say how far these two influences are responsible for the differences in his marks. By substituting these values in the left-hand side of our previous equation (vi), we can correct his marks, and virtually eliminate these two peculiar influences. In what follows I shall assume that these adjustments have been made; i.e., that each mark is expressed in "standard measure." Accordingly, let us write x for $\frac{X - \bar{X}}{\sigma_x}$ with a similar substitution for G.

[1] They cannot, of course, be ignored if we propose to compute differences between examiners which are of the same order, e.g., differences which themselves depend in part upon standard deviations. After all, "significance" is a relative term. Because a difference does not amount to 3 times its *p.e.*, that does not mean that it is devoid of *all* significance. In the absence of fuller material we are perfectly justified in making calculations from small groups, and even from technically "insignificant" figures, provided we state as precisely as possible how insignificant or how precarious the results may be.

[2] The first figure in either case is taken from Table 117, p. 187.

Section IV.—Measuring the Individual Examiner's Accuracy : the Coefficient of Correlation

566. We have now two sources of error left to contend with—S and E. The E's will be n in number, while the S's may be as many as $2^n - n - 2$. Now, we have seen that, as a rule, the influence of each S is comparatively slight in amount and comparatively narrow in range. With a well-chosen board of examiners, agreeing beforehand as nearly as possible on what qualities are relevant to the examination, the S's may be all but eliminated, or at any rate greatly reduced; what little is left can be treated as part of the E's—the effects tending to neutralize each other in different examiners.[1] And there are, as we shall discover in a moment, tests whereby in any particular case we can determine whether or not any specific factors, common to some examiners but not to all, have after all been operative to any discernible extent.

Treating the S's as negligible, and taking G and E to be expressed in standard measure, our equation is at last reduced to

$$x_k = b_{kg}\, g + b_{ke}\, e_k \qquad \ldots \text{(vii)}$$

The equation does not mean that if we knew a given candidate's true mark, g_j say, we could accordingly prophesy *exactly* what mark examiner k would give, but merely that we can calculate the mark he will *most probably* allot, i.e., what is the average of the marks that he would give to all the candidates who have g_j as their true mark. b_{kg}, therefore, is in effect what is sometimes called a regression coefficient. Nor can we, as a rule, use b_{kg} to deduce g from x. The customary notation is to use the symbol b_{kg} for the coefficient used in estimating x_k from g, and b_{gk} for the coefficient used in estimating g from x_k.

The geometrical mean of the two coefficients may be defined as the coefficient of correlation, and written

$$r_{kg}\ (= r_{gk}) \equiv \sqrt{b_{kg} \cdot b_{gk}}\ .$$

Like the cosine of an angle, a coefficient of correlation varies from -1 through 0 to $+1$; and measures, on that scale, the amount of concomitant variation between two variables. Thus r_{kg} indicates the agreement between the examiner's mark x_k and the ideal mark g; similarly, $r_{kk'}$ will indicate the amount of agreement between the marks from a pair of examiners, k and k'. I shall suggest in a moment that the coefficient r_{kg}, or some simple function of it, is the best single measure that we can obtain for

[1] For this reason I should prefer to speak of E as the *residual* variation rather than the *random* variation, so as to include the wider possibility that not all of it is due exclusively to random factors or blind chance.

measuring the efficiency of each examiner: it is, in fact, the weight that he implicitly attaches to the true mark. When the marks are in standard measure, $b_{kg} = b_{gk} = r_{kg}$: so that we may then write[1]

$$x_k = r_{kg}\, g + r_{ke}\, e_k \qquad \ldots \text{(viii)}$$

567. In what follows I assume that r is calculated by the method of product moments. The proof[2] may be summarized as follows. If there are N candidates, we have N equations of the type (viii) for determining b and therefore r. Accordingly, on applying the method of least squares, we obtain

$$r_{kk'} = \frac{\sum_{j=1}^{j=N} (x_{kj}\, x_{k'j})}{N\sigma_k\sigma_{k'}} \text{ or } \frac{1}{N}\sum_{j=1}^{j=N} (x_{kj}\, x_{k'j}) \qquad \ldots \text{(ix)}$$

if the marks are in standard measure.

Thus, to find $r_{kk'}$, we are in effect multiplying two rows of the third matrix in Table 134 (M_3)—the kth row and the k'th—and averaging the inner products. When we do this for the n rows of marks awarded by the n examiners, we obtain a new axisymmetric table (M_R), identical in form with our original matrix M_R (except for the margin of g coefficients, see p. 247) and consisting of the $\frac{1}{2}n(n-1)$ intercorrelations[3] of each examiner

[1] The assumptions embodied in this equation and the preceding are usually known as the two factor theory. Since $k = 1, 2, \ldots, n$, there will be $(n+1)$ independent co-ordinates indicated by g and e_k. Thus, if we take r_{kg} and r_{kc} to be direction cosines of the line x_k with reference to g and e_k, each equation of the type of (viii) above will describe a line that lies always in a two-dimensional plane. There are n lines of the type e_k. Hence, geometrically this means that the test-measurements are to be represented by points in a space of $(n+1)$ dimensions. Each plane passes through two of the $(n+1)$ axes; the $(n+1)$th axis is a common central axis representing the "general factor," and the remaining n axes are specific each to a single test. Every line or vector representing a test lies in one of these n orthogonal planes; none lies in any of the spaces between them.

The investigations on which this theory is based have mainly been concerned with tests of intelligence; but I have elsewhere shown (*Distribution and Relations of Educational Abilities*, pp. 52 *et seq.*) that the same appears to hold roughly true of examinations in elementary school subjects and also (in an unpublished research) of examiners as distinct from methods of examining. See also Note, p. 312.

[2] See any textbook of elementary statistics, e.g., Yule, *loc. cit.*, p. 161.

NOTE.—A rapid estimate of the correlation may be obtained by what is sometimes known as the footrule formula. If the two lists of marks are reduced to an order of merit, and if d_j = the difference between j's position in the first and second list, then the footrule coefficient

$$R = 1 - \frac{6\Sigma(d_j)}{N^2 - 1} \qquad \ldots \text{(ixa)}$$

With small groups, the calculation can be made in a few minutes. If desired, this rough coefficient may be approximately translated to the scale of the product-moment coefficient by taking $r = \sin\frac{\pi}{2}R$ or (if $R < \cdot 5$) $r = 1{\cdot}5\ R$.

[3] The table is usually printed with spaces for n^2 coefficients. But k's correlation with k' is the same as k''s with k; and the correlations of each examiner with himself are here ignored.

with his colleague. We thus reach a matrix equation which compactly describes the construction of a table of correlations from the mark-sheet, namely, $M_R = \frac{1}{N}(M_s M'_s)$ where $M_s \equiv$ matrix of marks in standard measure.[1] But x_{kj}, the mark awarded to any candidate j by examiner k, was, it will be remembered, originally considered equivalent to $\Sigma r_{ki}\, z_{ij}$, the product of the kth row of the first matrix (M_1) and the jth column of the second (M_2) in Table 134, i.e., $M_s = M_1 M_2$. Moreover, since the marks for the hypothetical elements are uncorrelated, $\Sigma z_{ij}\, z_{i'j} = 0$ $(i \neq i')$, i.e., $M_2 M'_2 = IN$. Thus we obtain a second matrix equation showing the theoretical relation of the table of correlations to the table of hypothetical weights, namely, $M_R = M_1 M'_1$; i.e., the whole table of intercorrelations is in theory to be obtained by post-multiplying the table of hypothetical weights by a transposed version of itself.

The significance of this important result will be clearer if we first reduce M_1 to fit the special case we are now considering: (see Table 137). Here only a weighting for the first element will be needed for all the examiners: each of the remaining elements will be peculiar to one examiner alone. The products are shown in the centre of the table. A later table (Table 140, p. 293) gives numerical coefficients obtained in this way.

[1] Incidentally the relation between M_R and M_S — the matrices given at the outset (p. 247) for determining the weights by the method of least squares—now becomes clear. Let M_X denote the matrix of candidates' marks, namely,

$$\begin{bmatrix} X_{g1} & X_{g2} \ldots X_{gN} \\ X_{11} & X_{12} \ldots X_{1N} \\ \ldots & \ldots \ldots \ldots \\ X_{n1} & X_{n2} \ldots X_{nN} \end{bmatrix}$$

and $\mathsf{T} = \mathsf{T}' =$ the diagonal matrix

$$\begin{bmatrix} \frac{1}{\sqrt{\Sigma(X_g^2)}} & 0 & \ldots & 0 \\ 0 & \frac{1}{\sqrt{\Sigma(X_1^2)}} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{1}{\sqrt{\Sigma(X_n^2)}} \end{bmatrix}$$

(The capital letter X indicates that the marks are not necessarily in standard measure, and the block letter M that hypothetical values for g are included). Then $\mathsf{M}_S = \mathsf{M}_X \mathsf{M}'_X$; and M_R, the theoretical matrix of correlations, $= \mathsf{T}\mathsf{M}_X \mathsf{M}'_X \mathsf{T}' = \mathsf{T}\,\mathsf{M}_S\,\mathsf{T}'$. If $\sigma_g = \sigma_1 = \ldots = \sigma_n = 1$, then $\mathsf{T}\,\mathsf{T}' = I\,N = N$ (where I is the unit matrix and N, as usual, the number of candidates. Consequently, with standard measure, $\mathsf{M}_R = \frac{1}{N}\mathsf{M}_S$ (the analogue of the formula given in the text). It will be observed that, if we took not σ but $\sigma\sqrt{N}$ as our unit, the scalar matrix would reduce to the unit matrix, and the operations be somewhat simplified. I may add that, to avoid excessively complicating an introductory account, I do not here distinguish between the *observed* values of the coefficients in M_R and the *theoretical* values (when the distinction seems necessary, I use $M_{kk'}$ for the latter); nor have I discussed what theoretical or observed values are to be entered for the diagonal of self-correlations. (On this point, see para. 579.)

TABLE 137

THE CORRELATIONS BETWEEN EXAMINERS AS THE PRODUCTS OF THEIR CORRELATIONS WITH THE GENERAL FACTOR

$M_{1'}$				$M'_{1'}$					
				Exr. 1	Exr. 2	...	Exr. n		
				r_{1g}	r_{2g}	...	r_{ng}		
				r_{1e}	0	...	0		
				0	r_{2e}	...	0		
				0	0	...	r_{ne}		
	General Factor	Individual Errors			$M_{R'}$			Totals	Products
Exr. 1	r_{1g}	r_{1e}	0	0	r_{1g}^2	$r_{1g}r_{2g}$ ...	$r_{1g}r_{ng}$	$r_{1g}\Sigma(r_{kg})$	$r_{1g}{}^n\Pi(r_{kg})$
Exr. 2	r_{2g}	0	r_{2e}	0	$r_{2g}r_{1g}$	r_{2g}^2 ...	$r_{2g}r_{ng}$	$r_{2g}\Sigma(r_{kg})$	$r_{2g}{}^n\Pi(r_{kg})$
...	...	...	...	...	...	... ...	...	...	...
Exr. n	r_{ng}	0	0	r_{ne}	$r_{ng}r_{1g}$	$r_{ng}r_{2g}$ ...	r_{ng}^2	$r_{ng}\Sigma(r_{kg})$	$r_{ng}{}^n\Pi(r_{kg})$
					Grand Total			$\Sigma(r_{kg})\Sigma(r_{kg})$	—
					Grand Product			—	$\Pi[r_{kg}{}^n\Pi(r_{kg})]$

More generally, if x_{jk} includes particular and individual factors as well as universal, the observed correlation $r_{kk'}$ becomes the sum of the products, two by two, of what we originally regarded as the weights for the separate elements, i.e.,

$$r_{kk'} = r_{k1}\,r_{k'1} + r_{k2}\,r_{k'2} + \ldots + r_{kq}\,r_{k'q} \qquad \ldots \quad (x)$$

or, if we group them as suggested above,

$$= \Sigma(r_{kg}\,r_{k'g}) + \Sigma(r_{ks}\,r_{k's}),$$

the signs of summation being inserted to allow for the possibility of more than one factor of each type. There are no terms $\Sigma(r_{ku}\,r_{k'u'})$ or $\Sigma(r_{ke}\,r_{k'e'})$ because k's weights for k''s "singular" elements and errors are by definition zero; and vice versa.

If, as before, we regard r_{kg}, r_{ks_1}, r_{ks_2}, etc., as direction cosines for the examiner k's mark-line OK (say), with analogous values for OK', then $r_{kk'}$ is determined by the familiar cosine law:—

$$\cos\theta = \cos\alpha\cos\alpha' + \cos\beta\cos\beta' + \cos\gamma\cos\gamma' + \ldots \quad (xi)$$

where $\theta \equiv KOK' = \cos^{-1} r_{kk'}$, and $\alpha, \beta, \gamma, = \cos^{-1} r_{kg}, \cos^{-1} r_{ks_1}, \cos^{-1} r_{ks_2}$, with corresponding values for α', β', γ', etc.[1]

[1] The formula is familiar to every student of geometry, and holds good no matter how many dimensions are involved (cf., e.g., Sommerville, *Geometry of N Dimensions*, p. 76). As applied to mental measurements this mode of formulation appears to have originated with Garnett (*loc. cit. sup.*, p. 96). Thurstone takes it as the starting point of his method of multiple factor analysis (*Psych. Rev.*, XXXVIII, p. 413).

568. The coefficient of correlation cannot be taken as measuring an examiner's efficiency on any absolute scale. r varies with the standard deviation of the group on which it is based. If a batch of candidates is perfectly homogeneous—all of identical merit—then r_{12} would approximate to zero, however accurately each examiner marks. Hence r is relative to the internal variation or variance (squared standard deviation) of the group that is measured.

As with the mark-sheet as a whole, so with the mark-list for each of the individual examiners: our aim is to analyse the total variance into the various contributory portions. Accordingly, if r is taken as meaning the relative weight attaching to the different factors, it can be conveniently thought of as expressing the ratio of the several variances. Thus, if we simply split the total variance of a given examiner's marks (σ_{kx}^2) into two elements—that due to the influence of the common element g and that due to irrelevant or erratic influences, r_{kg}^2 may be taken as measuring the proportion of the variance due to g, and $(1 - r_{kg}^2)$ the proportion due to all other factors. Thus

$$r_{kg}^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \qquad \ldots \text{(xii)}$$

And we may consequently write

$$x_{jk} = r_{kg}\, g_j + \sqrt{1 - r_{kg}^2}\; e_j \qquad \ldots \text{(xiii)}$$

$$= g_j \cos \alpha + e_j \sin \alpha \qquad \ldots \text{(xiv)}$$

i.e., x_{jk} is compounded according to the familiar vector law, and forms the sum of the projections of g_j and e_j (measured along co-ordinates at right angles) on the mark-line making an angle $\alpha = \cos^{-1} r_{kg}$ with the g-line. If we retain the other factors, then k's weighting for the general factor becomes

$$r_{kg} = \sqrt{\frac{\sigma_{kg}^2}{\sigma_{kg}^2 + \sigma_{ks}^2 + \sigma_{ku}^2 + \sigma_{ke}^2}} \qquad \ldots \text{(xv)}$$

$= \sigma_{kg}$, since we are taking his total variance as 1.

As before, his divergence from the general factor will be expressed by

$$\sqrt{1 - r_{kg}^2} = \sqrt{\sigma_{ks}^2 + \sigma_{ku}^2 + \sigma_{ke}^2} \qquad \ldots \text{(xvi)}$$

His divergence from everybody else—his "individuality," as it might be called, will be expressed by $\sqrt{\sigma_{ku}^2 + \sigma_{ke}^2}$. The measure of his self-consistency—his "reliability coefficient," as it is sometimes termed (r_{kk})—will be $\sigma_{kg}^2 + \sigma_{ks}^2 + \sigma_{ku}^2$;

so that his "unreliability," or inconsistency as it might better be termed, will be measured by

$$\sqrt{1 - r_{kk}} = \sqrt{\sigma_{ke}^2} = \sqrt{1 - r_{kg}^2 - r_{ks}^2 - r_{ku}^2} \quad \ldots \text{(xvii)}$$

If σ_{kx} is not unity, each of these ratios will have the same denominator as (xv).[1]

569. In psychological testing it has long been customary to measure the efficiency of a *test* by means of a coefficient of correlation. To measure the efficiency of a *person* by the same device is, however, a somewhat novel proposal. Ordinarily we take the figures obtained from persons to measure the efficiency of a test. Here we are in effect taking a figure obtained by means of a test to measure the efficiency of a person. Thus, instead of using the examiners to examine the candidates' scripts, we are using the candidates' scripts to examine the examiners. The change in standpoint involves several modifications in the customary procedure; but I hope to show that in principle, though not perhaps in detail, the statistical technique which has been worked out for factorizing the results of a test is equally valid for factorizing the operations of a person.[2]

[1] Reliability coefficients are in constant use to measure the self-consistency of mental tests. It is urgently to be desired that similar coefficients be collected for examiners. With a fairly wide interval of time between the two markings, I find such coefficients range usually from ·55 to ·95, averaging about ·85. They vary greatly for different subjects as well as for different individuals.

[2] Dr. Rhodes, in discussing the procedure I have suggested, has expressed a doubt whether correlation can validly be applied in this way. I may therefore add that I have already subjected this somewhat novel application to a practical test in several different fields. With the material analysed in the present book we have no means of assessing the efficiency of examiners apart from internal evidence. But with the marks from a University examination taken by nearly four hundred candidates (Teacher's Diploma, 1925–31) I have regularly employed the procedure here described on a fairly large scale with a view to measuring the apparent efficiency of the several examiners. There examiners and candidates were alike personally well known; and I could secure independent estimates to check my inferences. In earlier researches I had used the same procedure for analogous problems in vocational guidance. For example, in order to test efficiency in examining wool by touch, I asked wool-sorters to arrange fabrics in order of texture, heaviness, and the like. It was then possible to measure the tactile discrimination of each "examiner" by correlating his order with that of a weighted average furnished by all the testees or by specially chosen experts. In some cases, an "ideal" order could also be obtained from physical measurement (e.g., for weight, by actually weighing the fabrics; for texture, by actually measuring the threads); and it was thus possible to check the efficacy of the statistical device. (*J. Nat. Inst. Ind. Psych.*, I, 1922, p. 93, cf. *J. Text. Inst.*, Dec., 1926, p. 172. See also *A Study in Vocational Guidance*, 1926, p. 59 for other uses of the correlation between persons.) In a later publication I have reported experiments showing that the same inversion of the ordinary correlational analysis was practicable for estimating such qualitative characteristics as artistic capacity and determining temperamental types.

[*Footnote continued on next page.*

570. The upshot so far of the foregoing argument amounts to this: that we may most conveniently measure the accuracy or efficiency of a given examiner by the use of a coefficient of correlation. This proposal might have seemed obvious at the start. In deducing it at length my purpose has been to bring out the several assumptions involved and the various influences that are successively treated as negligible.

The next problem is to find a standard to represent *g* with which each examiner's marks may be correlated. The first and the most natural suggestion will be to take the average of all the examiners and correlate the marks of each one with this. For purposes of illustration let us again turn to the table of marks which has been used in the text of this book. The correlations of each of the six examiners of Board 1 who marked the Latin papers in the School Certificate Examination are shown on the next page (Table 138).

It will be observed that the peculiarity of this approach lies in using a coefficient of correlation to measure the agreement of an individual's performance *with a standard*. Now all test-performances are really marked according to agreement with a standard—the objective physical measurement of each stimulus in a sensory test, the key containing the correct answers in a test of intelligence or scholastic knowledge. If, for example, a child of 10 answers 40 test-problems out of 100 when the norm at that age is 50, the result could be expressed quite as well by a correlation (e.g., by a tetrachoric coefficient) as by a "mental ratio" or "I.Q." Where, on the other hand, there is (as here) no external or objective criterion, the standard—that is the "true" set of marks that would be awarded by an ideal examiner—has to be deduced from the several mark-lists furnished by various persons themselves. That is the only obvious difference. Indeed, a recent writer has actually claimed that the "general factor" usually described as intelligence arises solely from the presence of this all-pervading standard. (H. F. Adams, "The Theory of Two Factors: an Alternative Explanation," *Journal of Applied Psychology*, XV, 1931, pp. 16–34, and 358–377.) I should accept this argument as applied to the "general factor" in our present problem, but not in the more familiar case of the interrelations of intellectual abilities. The difference is as instructive as the analogy. Adams, in my view, commits the fallacy of confusing correlations between tests with correlations between persons. He therefore assumes that the current formulæ for factor-analysis are applicable to both without modification. Although in this memorandum I have mainly stressed the similarities, I believe that much fruitful work remains to be done by investigating the differences between the two lines of approach. Professor Godfrey Thomson has pointed out (*Brit. Journ. Psych.* XXVI, 1935, pp. 75–76) that the chief theoretical difficulty lies in discovering a comparable unit. We can correlate measurements for two characteristics—say height and weight—obtained from 100 persons: can we correlate for two persons the measurements obtained for 100 characteristics? How can we apply a product-moment formula to 60, 40, 12, ... when 60 means 60 inches, 40 means 40 pounds, 12 means 12 mental years, and so on? Similarly, how can we assume that 60, 40, 12 ... are in the same units when 60 is Examiner A's mark, 40 B's, and 12 C's? My answer would be that we must either use covariance, or else first reduce the crude scores to terms of the same unit, e.g., the standard deviation or (for rougher purposes) ranks in order of merit (cf. *sup.*, footnote 1, p. 253). It will be observed that this is the principle I am adopting here; and I should be tempted to criticize the method followed in the body of this book (Part II, pp. 186 *et seq.*), were it suggested for universal application, on the ground that each examiner's marks have not first of all been reduced to terms of his own standard deviation.

One advantage of using the coefficient of correlation is that we can easily make an estimate of the significance or "probable error" of the figure so reached. Let me repeat that in investigations such as the present this is to my mind essential. It might be positively misleading to state that C's marking correlates with the average to the extent of ·78 and with D's to the extent of ·47, if with groups so small as these such figures might easily have arisen by sheer chance.[1]

Here, whatever test is applied, every one of the coefficients in the table proves large enough to be fully significant statistically.

TABLE 138

MARKS FOR SCHOOL CERTIFICATE, LATIN: GROUP I

CORRELATIONS OF THE SEVERAL EXAMINERS' MARKS WITH THE AVERAGE OF THE MARKS AWARDED BY ALL

Examiner	A	B	C	D	E	F	Average
Correlation	·926	·924	·778	·891	·889	·943	·891
Probable Error	·026	·026	·071	·037	·038	·020	·037

571. To those who are unfamiliar with the implications of such coefficients, the correlations may at first sight seem high: their average amounts to nearly ·9. The implications may perhaps best be exhibited by employing a further coefficient to express the reduction of error. This can be calculated by the simple formula $\sigma_e = \sqrt{1 - r^2}$ (cf. equation (xvi), p. 274 above). Such a coefficient has been termed by French writers a "*coefficient de dispersion liée*," and by American a "coefficient of alienation."[2] Since it is complementary to the coefficient of correlation, I

[1] In Table 138 I give probable errors calculated by the familiar formula $\cdot 6745 \times \frac{1 - r^2}{N - 1}$, where $N \equiv$ the number in the group. A more accurate test is to take

$$z = \log_e \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad \text{or} \quad t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

and to employ the tables showing the distribution of z or t (Fisher, *loc. cit.*, pp. 232 or 158). Enter the tables for z with $n_1 = 1$ and $n_2 = N - 2$ or t with $N - 2$. But when, as here, the problem is simply to decide whether any correlation is present or not—i.e., whether the ascertained figure differs significantly from zero—a rough and rapid criterion is to require that the coefficient should at least equal $2 \times \frac{1}{\sqrt{N}}$: thus, if there are, say, only 15 or 16 candidates in the group, the smallest significant coefficient will be roughly $2 \times \frac{1}{\sqrt{16}} = \cdot 5$.

[2] The latter term is Kelley's (*loc. cit.*, p. 173 and ref.); but neither term has come into general use.

should prefer to call it the " coefficient of non-relation."[1] Thus, if r (the measure of correlation) = ·89 (as for the average), $\sqrt{1 - r^2}$ (the measure of non-relation) = ·46, or nearly one-half. Again, if r = ·78 (as for C), $\sqrt{1 - r^2}$ = ·61, or nearly two-thirds. The fraction so obtained shows the extent by which an error due to a sheer chance judgment is diminished in virtue of the influences measured by the correlation. It gives in fact the amount of residual error, when the marks are translated into terms of the examiner's standard deviation taken as unity.

Thus, if Examiner C had awarded each of his marks entirely at random, e.g., by simply drawing figures from a hat, his probable error would be about 2·6 marks : (with a group of this size the probable error of the individual mark is roughly one-fifth or one-sixth of the examiner's total range). Having read the scripts, C has reduced this amount of error by about two-thirds, that is to about 1·7 marks.[2] The other examiners have succeeded in about halving the errors they would have made, had they marked blindly and at random. Even with these high correlations, therefore, a comparatively small amount of accuracy is achieved. For the other examination results analysed in this book, where the correlations are much smaller, the amount of error remaining must be astonishingly large.

572. Between the size of the individual correlations there is no great difference. Accordingly, before discussing whether any particular examiner is more accurate than his colleagues, it will be essential to apply some valid test to determine what differences, if any, are significant. There is no justice in stating that the validity of C's marking is only four-fifths of that of F, if after all the difference between the calculated coefficients might easily have arisen from the chances of random sampling.

[1] The non-relationship can be given a positive interpretation : if x correlates with a fundamental factor g to the extent of r_{xg} and if there exists some other factor e which is independent of g, but which together with g completely determines x, then the correlation of x and e is $\sqrt{1 - r_{xg}^2}$. Or, reverting to the twofold analysis of variance suggested above (p. 274), it follows that, if

$$r_{kg}^2 = \frac{\sigma_{kg}^2}{\sigma_{kg}^2 + \sigma_{ke}^2}, \text{ then } 1 - r_{kg}^2 = \frac{\sigma_{ke}^2}{\sigma_{kg}^2 + \sigma_{ke}^2},$$

and so measures the ratio of the *residual* variance to the *total* variance. $\sqrt{1 - r_{kg}^2}$ in fact is a standard error of estimate as expressed in standard measure.

Multiplying this coefficient by the observed standard deviation, we obtain the familiar " standard error of estimate " expressed in terms of the original measurements or marks—i.e., the error made in estimating x_k from g, ($\sigma_{kx}\sqrt{1 - r_{kg}^2}$). This, or rather its square, is the quantity given above in Table 117 (p. 187), as the (residual) " variance " (cf. also Table 142, p. 294 below).

[2] His standard error is 3·9 and he reduces it to about 2·7 : see below, Table 142.

Calculated by the usual formula[1] the probable error of the difference is ± ·074. The difference itself (·943 — ·778 = ·165) is 2·24 times this probable error. Such a difference would occur by chance about once in 7 or 8 times—that is to say, in comparing half a dozen coefficients like those in the table (15 comparisons in all) it would actually be surprising if there were no such difference. Since then all the examiners correlate with the average of the whole board to much the same extent, it would seem to follow that, in determining the "ideal" mark to be allotted to the candidates, we might safely take the simple and unweighted average as indicating the ideal (as, for example, has been done by way of a first approximation in the body of this book in para. 394).

In other examinations, a very different situation may be found. In my earlier analysis of the results of a University examination, covering nearly 400 candidates, I found that the correlation between examiners marking the same scripts—an internal and an external examiner, for example—might occasionally sink as low as ·4 or even lower. (Low coefficients occur still more frequently when the different examiners are marking the same candidates for *different* branches of the general subject.) Thus, once again the problem arises: how can we determine the "true" marks for the candidates, when examiners differ so widely in the accuracy of their assessments? No longer will it seem legitimate simply to average each examiner's marks just as they stand. Even with School Certificate Latin the method is obviously of doubtful validity in the case of C. When correlating C's marks with the average we have included in that average C's own marks; this must plainly exaggerate his apparent agreement with the ideal. Had we omitted his mark from

[1] See footnote 1, p. 277. Where the coefficients are high, and particularly where the groups are small, the familiar criterion based on the formula used above is apt to be inexact. In such cases the distribution of the coefficients is not normal. The simplest plan is to compare the values, not of r but of $z = \tanh^{-1} r$, for which the distribution is very nearly normal (cf. Fisher, *Statistical Methods for Research Workers*, p. 185 *et seq.*; Dawson, *Computation of Statistics*, p. 140: both give tables; or the tables of natural logarithms or hyperbolic functions may be used). These values are given in Table 139 below. Then, in considering the significance of the differences, we may take the probable error of each coefficient as approximately $\frac{\cdot 6745}{\sqrt{N-3}}$, i.e., here ± ·195. It is independent of the value of the correlation; consequently the probable error of the difference between any two of the values of z will be ± ·276. No difference reaches three times this value; hence none is significant.

TABLE 139

SCHOOL CERTIFICATE LATIN: GROUP I

VALUES OF $\tanh^{-1} r$ FOR THE SEVERAL EXAMINERS

Examiner	A	B	C	D	E	F	Average
$\tanh^{-1} r$ . .	1·63	1·62	1·04	1·43	1·42	1·76	1·43

our estimate of the ideal, his correlation would have been significantly lower; still, to omit his marks altogether seems almost as unjustifiable as to include him on equal terms with the rest.

573. Evidently a better estimate of the ideal mark could be obtained by keeping each examiner's contribution, but weighting his marks according to his relative accuracy. How, then, are we to determine the weights? As a first approximation we might begin by using for our weights the correlation of each examiner with the unweighted average. This weighted average would then yield a slightly better approximation to the "ideal" mark. We could thus recalculate both the correlations and the weighted average; and then repeat the process until no appreciable change in the coefficients was produced by the further recalculations.

To build up these successive approximations, however, would often prove a very laborious process; and the assumptions we have already made will enable us to determine the correlation of each examiner with the hypothetical true marks more directly by a perfectly simple formula. Let us, therefore, proceed to inquire how such a formula may be deduced.

Section V.—To Determine the Correlation between the Marks of a Given Examiner and the Hypothetical True Marks: the "H.G.F." or "Saturation Coefficient"

574. By multiplying the matrix of weightings by the transposed version of itself we found

$$r_{kk'} = r_{k1}\, r_{k'1} + r_{k2}\, r_{k'2} + \ldots + r_{kq}\, r_{k'q} \qquad \ldots \text{(xviii)}$$

(equation (x), p. 273), where r_{k1}, $r_{k'1}$, ... denote the weights attached by examiners k and k' to the 1st, 2nd, ... qth elements respectively. If only the 1st element is common to all the examiners and is consequently identifiable with g, the equation reduces to

$$r_{kk'} = r_{kg}\, r_{k'g}$$

where r_{kg} denotes k's correlation with the hypothetical general factor g (his saturation or h.g.f. coefficient, as it is sometimes termed).

This result may most simply be proved without the complications of matrix multiplication as follows.

575. Let $x_{1j}, x_{2j}, \ldots, x_{kj}, \ldots, x_{nj}$ be the marks awarded by a series of n examiners to any given individual j. Let the marks be measured as deviations from the average of each examiner, and let their standard deviations be $\sigma_1, \sigma_2, \ldots, \sigma_k, \ldots, \sigma_n$. Using the same notation as before, $b_{kg} \equiv r_{kg} \frac{\sigma_k}{\sigma_g}$ will represent the weight

which k gives to g, i.e., the so-called regression coefficient for determining the most probable value of x_{1j} for any given g_j. If the marks are the result of components of two kinds, first the hypothetical general factor influencing all the examiners but in different degrees, and, secondly, a series of n independent components peculiar to each examiner (his error, if we please so to regard it), we may put

$$x_{1j} = b_{1g} \cdot g_j + e_1'$$
$$x_{2j} = b_{2g} \cdot g_j + e_2'$$
$$\ldots \qquad \ldots\ldots\ldots \qquad \text{etc.}$$

Thus, the correlation between x_{1j} and x_{2j} (i.e. between the mark-lists of the two examiners, 1 and 2),

$$r_{12} = \frac{\Sigma(x_{1j}\, x_{2j})}{n\ \sigma_1\ \sigma_2} \qquad (j = 1, 2, \ldots N)$$
$$= \frac{\Sigma\ \{(b_{1g} \cdot g + e_1)\ (b_{2g} \cdot g + e_2)\ \}}{n\ \sigma_{x1}\ \sigma_{x2}}$$
$$= \frac{b_{1g}\, b_{2g}\, \sigma_g^2}{\sigma_1\ \sigma_2}$$

(the other terms in the product disappearing, because r_{ge_1}, r_{ge_2}, and $r_{e_1e_2}$ all $= 0$)

$$= r_{1g}\, r_{2g}.$$

And generally

$$r_{kk'} = r_{kg}\, r_{k'g} \qquad \ldots \text{(xix)}$$

In other words, if there is no "specific influence" common to the two examiners, then, theoretically, the correlation between k and k' should be the product of their two respective correlations with the true marks.

576. From this two corollaries follow :—

First, $r_{12}\, r_{34} = r_{1g}\, r_{2g}\, r_{3g}\, r_{4g}$
$r_{13}\, r_{24} = r_{1g}\, r_{2g}\, r_{3g}\, r_{4g}$

Hence, $r_{12}\, r_{34} - r_{13}\, r_{24} = 0 \qquad \ldots \text{(xx)}$

or $\dfrac{r_{12}}{r_{13}} = \dfrac{r_{24}}{r_{34}} \qquad \ldots \text{(xxi)}$

The left-hand side of equation (xx)—sometimes known as a "tetrad difference"—represents a minor determinant of the second order from the original matrix of intercorrelations. When all such 2×2 minors vanish, all the intercorrelations can be explained by postulating one all-pervading factor and one only : this will be obvious on considering the 2×2 minors of M_3' in

Table 137 (p. 273, above). Such a matrix is said to be of rank 1. On the other hand, if they do not vanish, or at least approximate to zero within the limits indicated by their probable errors, the present method of analysis can lead only to provisional or incomplete conclusions. But, on considering the origin of the terms in equation (xviii), it becomes clear that the theorem just enunciated can be given a more general form : the smallest number of factors that will account for a set of intercorrelations is indicated by the rank[1] of the set. When equations (xx) and (xxi) are satisfied throughout the whole set of coefficients, i.e., when any two rows or columns are proportional, the coefficients are said to form a " hierarchy."[2]

Secondly, $$\frac{r_{12}\, r_{13}}{r_{23}} = \frac{r_{1g}\, r_{2g}\, r_{1g}\, r_{3g}}{r_{2g}\, r_{3g}} = r_{1g}^2 \qquad \ldots \text{(xxii)}$$

Consequently, unless all three coefficients are positive or only one coefficient is positive, r_{1g} (and consequently r_{2g} and r_{3g}) will be imaginary ; and again the simpler method of analysis will be inapplicable.

The obvious causes that might lead to an infringement of these two conditions are (i) the presence of " specific " influences, common to two (or more) examiners but not to all ; (ii) the presence of large fluctuations in the observed correlations arising from the small size of the group examined—i.e., from the errors of sampling.

577. We have, then, as an equation for determining the

[1] The " order " of a determinant (or matrix) is defined by the number of rows (and/or columns) ; its " rank " by the highest order of minors that do not vanish. In earlier books on algebra " rank " is sometimes used for " order."

[2] Equation (xxi) was first given in that form in my early article on intelligence-tests in *Brit. Journ. Psych.*, III, 1909, p. 159 : but, as is there pointed out, it is immediately deducible from Prof. Spearman's previous work (*Zeitschr. f. Psych.*, XLIV, 1906, p. 85 ; cf. also *Amer. Journ. Psych.*, XV, 1904, pp. 274–5). Statistical readers will see that it follows at once from Udny Yule's well-known formula for partial correlation (*Introduction to Statistics*, p. 239 : see below, equation (xliv)). It serves as the most convenient preliminary criterion for determining by rough inspection the tendency towards hierarchical arrangement. For a more exact check Prof. Spearman has devised an ingenious method of procedure which depends on comparing all the tetrad-differences given by equation (xx) with their probable error : see *Abilities of Man*, Appendix, pp. x *et seq.* As regards the application of the general factor theory to the problem of intelligence-testing, Prof. Godfrey Thomson has advanced certain important criticisms (see Brown and Thomson, *Essentials of Mental Measurement*, esp. ch. X, sect. 5, on " Hierarchical Order as the Natural Order among Correlation Coefficients " ; also Thomson's later investigations of the mathematical aspect of the problem in recent numbers of the *Brit. Journ. Psych.*, especially XXVI, 1935, pp. 63-92). Thomson's attitude in regard to the point at issue here may perhaps be summed up in his statement that if the tetrad-differences are zero " it is possible, though not imperative, to postulate a mathematical quantity g which forms the sole source of the correlations " (*loc. cit.*, XXV, 1934, p. 94) ; but this (as he has pointed out) is only a partial statement of his whole position. The application of the general factor theory to the present problem appears to Professor Thomson (as I understand from private discussion) to be perfectly valid.

correlation of any examiner (say No. 1) with the ideal mark

$$r_{1g} = \pm \sqrt{\frac{r_{1k}\, r_{1k'}}{r_{kk'}}} \qquad \ldots \text{(xxiii)}$$

where k and k' are any two other examiners. With this formula, when there are more than three examiners, we shall have as many determinations for r_{1g} as there are pairs of examiners. Owing to chance fluctuations, these determinations will not exactly agree. We must therefore find some method for extracting the most probable estimate or average. There are, as I have pointed out in an early publication dealing with this problem,[1] several expedients that may be adopted for this purpose; and more than one has in point of fact been used in researches on mental testing.

578. If we have reason to assume only one general factor, then we can take the true values for $r_{kk'}$ to be those given by $M_{R'}$ in Table 137 (p. 273, above). Accordingly, to find a general determination for r_{1g} (say), we may now follow either the principle of the geometric mean or that of the arithmetic mean. The former involves taking the product of all the possible values; the latter their sum. (The sums and products are shown in the right-hand columns of the table.)

(i) For the product method, we simply take the geometric means of all possible determinations. We have

$$r_{1g} = \pm \frac{\sqrt[n]{{r_{1g}}^n \,.\, \Pi(r_{kg})}}{\sqrt[n^2]{\Pi\left\{{r_{kg}}^n \Pi(r_{kg})\right\}}} = \pm \frac{\sqrt{\Pi(r_{1k})}}{\sqrt[n^2]{\Pi(r_{kk'})}} \qquad \ldots \text{(xxiv)}$$

$$= \frac{\text{G.M. of Coeffs. in 1st Row}}{\text{G.M. of all Coeffs. in Table}}$$

(ii) For the summation method, we might simply take the arithmetic means of all possible determinations. In practice difficulties may be encountered with both methods.[2]

[1] *Brit. Journ. Psych., loc. cit. sup.*, p. 163.

[2] (1) $r_{kk'}$ may at times be nearly zero; consequently the resulting ratio obtained in calculating r_{jg} may be indefinitely large (the true value of r_{jg}, of course, cannot exceed unity). The presence of this indefinitely large value may greatly falsify the mean obtained by either multiplying or summing the several ratios.

(2) We might overcome this difficulty by considering the probable error of each observed coefficient: we could then weight the determinations accordingly, or simply ignore those based on values that are of no statistical significance. In any case, however, the labour involved will generally be out of all proportion to the gain in accuracy: for each examiner or test we may have to average $\frac{1}{2}n(n-1)$ quantities, that is $\frac{1}{2}n^2(n-1)$ for the whole table—nearly 5,000 with a set of 10 tests.

(3) The distributions tacitly assumed by adopting either the geometric or the arithmetic method of averaging are neither of them in strict conformity with what is known of the chance distributions of correlational coefficients: we might attempt to overcome this by working with the hyperbolic tangents; but the labour would be still further augmented, even if the procedure were shown to be valid.

By summing the numerators and denominators separately, however, we may in part evade these difficulties ; and so reach a slightly different formula, still analogous in its structure to that reached by the product method.

Taking an arithmetic mean in the ordinary way we have

$$r_{1g}^2 = \frac{1}{n}\left\{\frac{r_{11}r_{12}}{r_{12}} + \ldots + \frac{r_{12}r_{13}}{r_{23}} + \ldots + \frac{r_{13}r_{14}}{r_{34}} + \ldots\right\}$$

$$= \frac{r_{11}r_{12} + \ldots + r_{12}r_{13} + \ldots + r_{13}r_{14} + \ldots}{r_{12} + \ldots + r_{23} + \ldots + r_{34} + \ldots}$$

provided $\frac{r_{11}r_{12}}{r_{12}} = \frac{r_{12}r_{13}}{r_{23}} = \frac{r_{13}r_{14}}{r_{34}} = \ldots$ *exactly.*

On this assumption we may deduce a convenient formula as follows (I write it in a form which brings out the analogy with the result of the product-method) :—

$$r_{1g} = \pm \frac{\frac{1}{n}\left\{r_{1g}\,\Sigma\, r_{kg}\right\}}{\sqrt{\frac{1}{n^2}\left\{\Sigma\, r_{kg}\,\Sigma\, r_{kg}\right\}}} = \pm \frac{\Sigma(r_{1k})}{\sqrt{\Sigma(r_{kk'})}} \quad \ldots \text{(xxv)}$$

$$= \frac{\text{A.M. (or Sum) of Coeffs. in 1st Row}}{\text{Root of A.M. (or Sum) of all Coeffs. in Table}}$$

This simple formula, which I gave and used in an early publication,[1] is the speediest of all, and seems by far the best for most purposes where a quick approximation is required. It will be instructive to observe what it implies when the precise conditions under which it has been deduced do not hold good.

We have been assuming that all the intercorrelations are due to a single general factor only. Let us now suppose that there may be in addition other factors, common to some or all of the examiners and having a different scheme of weighting—specific factors, in short, such as those indicated in the matrices of hypothetical weightings in Tables 135 and 137 (pp. 264, 273 above). On determining $r_{kk'}$ by Garnett's "cosine law" (equation

[1] *Distribution and Relations of Educational Abilities* (1917), p. 53.

Originally I proposed to use the geometric mean, not so much because the quantity to be calculated has the form of a ratio, as because the manner in which the variations were distributed took a markedly asymmetrical shape. At that date, however, the distribution of correlation coefficients had not been investigated with exactitude (see *Biometrika*, Vol. XII, p. 125) ; further experience of such distributions led me later to prefer the formulæ given by (xxv) or (xxviii). There is evidently room for a more intensive study of the mathematical nature of the problem from the standpoint of the chance-distribution of the coefficients to be calculated ; but to determine along these lines what method of averaging is theoretically most appropriate would evidently involve a somewhat complicated piece of work.

xviii), and adding the totals and grand totals as before, we find :—

$$\frac{\Sigma(r_{1k})}{\sqrt{\Sigma(r_{kk'})}} = \frac{r_{1g}\,\Sigma\, r_{kg} + r_{1s_1}\,\Sigma\, r_{ks_1} + r_{ks_2}\,\Sigma\, r_{ks_2} + \dots}{(\Sigma\, r_{1g})^2 + (\Sigma\, r_{1s_1})^2 + (\Sigma\, r_{1s_2})^2 + \dots}$$

where s_1, s_2, ... denote the additional "specific" factors. If, therefore, we take the value given by equation (xxv) to represent the value of r_{1g}, we are implying either r_{ks_1}, r_{ks_2}, ... all $= 0$ (the assumption from which (xxv) was actually deduced) or else that the sums $\Sigma(r_{ks_1})$, $\Sigma(r_{ks_2})$, ..., each $= 0$, i.e., that r_{ks_1} (and r_{ks_2}) includes negative weightings as well as positive, and that the sum of the negative values (as with the deviations about a mean) equals that of the positive.[1]

579. In applying equation (xxv) as it stands there is what may seem at first sight a practical difficulty. Our tables of intercorrelations include no coefficient which can fairly represent the correlation of each examiner with himself. Even if we had them, the reliability coefficients would not serve: for they include, as we have seen, besides the influence of the general factor (and the overlapping specific factors, if any) the effect of the individual factors peculiar to the examiner himself, and their magnitude is thus unduly augmented. The difficulty can easily be overcome by successive approximation. In the spaces for these self-correlations trial values are inserted by smoothing the several columns. These are checked by computing r_{kg}^2 when r_{kg} has been found by equation (xxv). Unless the table deviates widely from hierarchical arrangement, a re-calculation of r_{kg} is seldom found necessary.

We can, however, if we desire, deduce a direct algebraic formula. By dropping the self-correlations from our theoretical table, we reduce the number of coefficients in each row from n to $n - 1$ and the number in the whole table from n^2 to $\frac{n(n-1)}{2}$ (since $r_{kk'} = r_{k'k}$).

Accordingly, (i) by the product method we have, after cancelling common terms in numerator and denominator,

$$r_{1g} = \pm \frac{\left\{\Pi(r_{1k})\right\}^{\frac{1}{n-2}}}{\left\{\Pi(r_{kk'})\right\}^{\frac{1}{2(n-1)(n-2)}}} \qquad \dots \text{(xxvi)}$$

[1] On referring to cases where such values have been calculated (the fullest collection are those given in Mr. Alexander's tables, *loc. cit. sup.*, pp. 33 *et seq.*) it will be seen that this is approximately so: the slight divergences are traceable to the special treatment of the self-correlations.

Thus the geometric mean is still available, and can be quickly computed by logarithms: but, as before, if the coefficients are low, it is apt to give impossible results.

(ii) For the modified summation method, we have

$$r_{1g}^2 = \frac{2}{(n-1)(n-2)} \left\{ \frac{r_{12}\, r_{13}}{r_{23}} + \frac{r_{12}\, r_{14}}{r_{24}} + \dots \right\}$$

And, proceeding as before, if each of the ratios within brackets were exactly equal, we could write

$$r_{1g}^2 = \frac{r_{12}\, r_{13} + r_{12}\, r_{14} + \dots}{r_{23} + r_{24} + \dots}, \text{ that is } \frac{\Sigma(r_{1k}\, r_{1k'})}{\Sigma(r_{kk'})}$$

(where $k \neq k'$ or 1; and $r_{kk'}$ and $r_{k'k}$, being equal, are not both included)

$$\text{or } r_{1g} = \pm \sqrt{\frac{\Sigma(r_{1k})^2 - \Sigma(r_{1k}^2)}{\Sigma(r_{kk'}) - 2\Sigma(r_{1k})}} \qquad \dots \text{ (xxvii)}$$

where $\Sigma(r_{1k})$ indicates as before the sum of all the coefficients in a row and $\Sigma(r_{kk'})$ the sum of all the coefficients in the table. This last formula is due to Professor Spearman[1]; and r_{1g}, as thus calculated, has been termed the "saturation-coefficient."

The foregoing proof, however, is open to a theoretical objection. Strictly it is invalidated by the very conditions it is framed to meet. Unless the ratios it seeks to average are unequal, there is no need for the formula; but if the ratios are unequal, the separate summation of numerators and denominators is no longer legitimate.

There is an obvious way of meeting the difficulty. Including self-correlations, we have $\frac{1}{2}n(n+1)$ equations to determine n quantities; and the several equations are likely to be inconsistent.

(iii) Accordingly, following the method of least squares, let us put

$$r_{jk} - r_{jg}\, r_{kg} = e_{jk} \qquad \begin{matrix} k = 1, 2, \dots, n \\ j = 1, 2, \dots, n \end{matrix}$$

where e_{jk} stands for the errors or residuals. In accordance with the usual procedure, square the expression on the left, sum first for $k = 1, 2, \dots, n$, and then for $j = 1, 2, \dots, n$, in turn.

[1] Professor Spearman's statistical methods have been developed in a long series of original papers by him, beginning with the early and supremely important contribution in the *Amer. Journ. of Psych.*, XV, 1904, pp. 268 *et seq.* They are concisely summarized in the Appendix already cited (*The Abilities of Man* (1927), pp. i–xxii).

Differentiate in respect of r_{jg}, and set the first derivative = 0; we finally reach

$$-\Sigma r_{kg} r_{jk} + r_{jg} \Sigma r_{kg}^2 = 0$$

$$\text{i.e.,} \quad r_{jg} = \frac{\Sigma(r_{kg} r_{jk})}{\Sigma(r_{kg}^2)} \qquad \ldots \text{(xxviii)}$$

$$= r_{1g} \frac{\Sigma(r_{kg} r_{jk})}{\Sigma(r_{kg} r_{1k})} \qquad \ldots \text{(xxix)}$$

This equation, as we shall find in a moment, yields a decidedly closer fit to the actual tables to which it is applied[1]; and, with certain obvious modifications necessitated by slight changes in the initial assumptions, the simple proof just given leads to a formula identical with that derived by Hotelling for his first "principal component."[2]

580. It will be observed that the various formulæ for finding r_{kg} given by equations (xxv) to (xxix) are all in their essence approximation formulæ. Their apparent divergence, therefore,

[1] In applying this formula, as with the preceding, we start by taking trial-values for each r_{jg}, and assume each $r_{jj} = r_{jg}^2$; we then calculate the right-hand side of equation (xxix), and observe whether the result agrees with the original figure. For the final value use equation (xxviii).

In its general form the equation reached above is analogous to the preceding (xxv). The differences are similar to those between a standard deviation and an arithmetic mean. The present formula may be deduced from the identity

$$r_{jg} \equiv r_{jg} \frac{\frac{1}{2n}\sqrt{2\Sigma(r_{kg}^2)}}{\frac{1}{2n}\sqrt{2\Sigma(r_{kg}^2)}}$$

just as the other (xxv) is deduced from the identity

$$r_{jg} \equiv r_{jg} \frac{\frac{1}{n}\Sigma(r_{kg})}{\frac{1}{n}\sqrt{\{\Sigma(r_{kg})\}^2}}$$

The principle implicit in both is an endeavour to equate the scales of the observed coefficients and the theoretical by calculating moments. It will be remarked, however, if we convert (xxix) into standard deviation form, we take 0 as the origin (hence the $2n$ values): but this assumes that negative values of r_{kg} are as likely as positive, which is not the case.

If, accepting (xxv) and taking $x_{kj} = r_{kg} x_{gj} + r_{ks} e_j$, we sum for k, we have apparently $\Sigma(x_{kj}) = x_{gj} \Sigma(r_{kg}) + 0 = x_{gj} \sqrt{\Sigma(r_{kk'})}$, or $x_{gj} = \Sigma(x_{kj})/\sqrt{\Sigma(r_{kk'})}$. Thus, x_{gj} (j's "true mark" as thus determined) has the appearance of being simply an unweighted average of the original marks (or rather their sum expressed as a multiple of the *s. d.* of such sums). But the assumption $(\Sigma r_{ks})^2 = \Sigma r_{kk'} - (\Sigma r_{kg})^2 = 0$ exactly (though made by Thurstone: cf. p. 306) is scarcely valid. For a better mode of determination, see p. 297 *et seq.*

[2] *Journ. Educ. Psych., loc. cit. sup.*, p. 429. Hotelling works with ratios instead of with the h.g.f. coefficients themselves; and, presumably because he assumes no co-ordinates for sampling errors, he starts with "corrected" coefficients and puts the self-correlations = 1·00. In practice this latter assumption seems peculiarly apt to disturb the relations of "general" and "specific" factors.

need trouble us no more than the divergences commonly found in other instances between alternative methods of averaging. Which formula we prefer will depend mainly on the underlying conceptions with which we are working.

(i) If we adopt what I have called the method of F-axes we shall incline rather to the simple summation-formula. Having n tests or examiners whose standard deviations have been equalized, we shall represent their test-lines by vectors in n-dimensional space. Then, to obtain a common component which shall represent them all, it will be natural to average their deviations by taking first moments (equation (xxv)). This procedure is equivalent to what Thurstone has recently called the centre of gravity method.[1] I used it in my earlier researches, and, as noted above, found it the best for first approximations.

(ii) If there is evidence that one factor is of overwhelming influence, while the influence of the others is narrow and slight, we shall seek a common axis where all the test planes intersect. This is the method of Spearman and his followers: since they omit the reliability coefficients from their matrix, they have usually adopted the summation formula in the shape given by equation (xxvii).

(iii) The product-method (xxvi) has recently been suggested by Thurstone as "applicable to the special case of the Spearman g-factor problem,"[2] and as therefore forming an alternative to Spearman's equation (xxvii).[3] After my first investigation on the so-called g-factor I rejected it on empirical grounds as giving the poorest fit of all the methods then tried; and it seems hardly in conformity with what is now known of the chance distribution of r.

(iv) If we start with what I have called T-axes, then, just as in dealing with two tests only, we take the principal axis of the frequency ellipse to represent the factor that accounts for the greatest amount of variance, so in dealing with n tests or n examiners we may take as representing the chief common factors the principal axes of the n-dimensional ellipsoid; and the most general factor of all will be that which accounts for the greatest amount of total variance and will therefore be represented by the longest axis. This in turn means finding a factor, g, say, which shall be such that the sum of the squares of all correlations like r_{kg} shall be a maximum; and leads to a formula analogous to equation (xxviii).

[1] *Psych. Rev., loc. cit. sup.*, p. 414.

[2] *Theory of Multiple Factors*, 1933, p. 62.

[3] It has also been used by Dr. Rhodes in applying what he terms the "new method" to the present problem (see Appendix II, para. 613).

581. If we envisage the possibility that there may be more than one general factor, the problem becomes more complex. The simplest approach would be to follow the procedure adopted in measuring the correlational angle where two variables only are concerned. There the two axes of reference are rotated until they coincide with the major and minor axes of the contour ellipses, with the result that the coefficient of the product-term ($2r_{12}$) is eliminated from the equation for the ellipses.[1] The proof usually adopted for two variables, however, cannot be transferred as it stands to the case of three or more variables; but the procedure applicable to the latter case can always be used for the simpler. It may be summarized as follows:—

Let $r^{kk'} \equiv \frac{\Delta_{kk'}}{\Delta_R}$, where Δ_R is the determinant formed from the matrix of correlation coefficients M_R, and $\Delta_{kk'}$ is the complementary minor of $r_{kk'}$ in Δ_R. Then, by the theory of multiple correlation, the ellipsoids of uniform frequency will be given by

$$F(x_1\, x_2, \ldots, x_n) \equiv F \text{ (say)} \equiv \sum_k^n \sum_{k'}^n (r^{kk'}\, x_k\, x_{k'}).$$

For purposes of matrix manipulation, this equation may be written more concisely $X'MX$, where X' denotes the one-rowed matrix $(x_1, x_2, \ldots x)$ and M denotes the matrix of the quadratic form; i.e.,

$$M \equiv \begin{bmatrix} r^{11} & r^{12} & \ldots & r^{1n} \\ r^{21} & r^{22} & \ldots & r^{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r^{n1} & r^{n2} & \ldots & r^{nn} \end{bmatrix}$$

Since $r^{kk'} = \frac{\Delta_{kk'}}{\Delta_R}$, $M = M_R{}^{-1}$. To determine the principal axes of the ellipsoid, the original rectangular axes of reference must be successively rotated until they coincide with the principal axes of the ellipsoid. Accordingly let L denote the matrix of direction cosines, and ρ the radius vector. In polar form the equation may be written $L'\, M\, L = \frac{C}{\rho^2}$. The problem then is to determine the maximum value of ρ under the condition that

[1] The two correlated variables are thus reduced to terms of two uncorrelated variables, whose variance is respectively a maximum and a minimum. See above, p. 254, and, for the usual proof, Yule, *loc. cit.*, p. 321.

$L'\,L = 1$. As in the case of two variables, this is equivalent to eliminating the product terms from F and so reducing it to a sum of squares, such as

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 + \ldots + \lambda_n x_n^2$$

The usual procedure is as follows. If the discriminant of F is not zero, it is always possible to find real values of a multiplier λ (say) such that the discriminant of $F - \lambda(x_1^2 + x_2^2 + \ldots + x_n^2)$ shall be zero; and it can then be shown[1] that the coefficients λ_1, λ_2, λ_3, ... are identical with the roots of the equation $|M - \lambda I| = 0$.

Multiply both sides of this equation by M^{-1} (that is by the matrix of observed coefficients, M_R); expand; divide each row by $-\lambda$. We thus reach the following equation for determining $\frac{1}{\lambda}$:—

$$\begin{vmatrix} r_{11} - \frac{1}{\lambda} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} - \frac{1}{\lambda} & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & \ldots & r_{nn} - \frac{1}{\lambda} \end{vmatrix} = 0$$

It follows that the roots of this equation, $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \frac{1}{\lambda_3}, \ldots$ are proportional to the lengths of the principal axis. The longest axis ($\frac{1}{\lambda_1}$ say) will be that of the most general factor of all (g say),

[1] The proof, originally due to Cauchy (*Œuvres*, 1829, 2me sér., Vol. IX, pp. 175 *et seq.*) is to be found, with various modifications, in most textbooks on solid geometry (e.g., W. H. Macaulay, *Solid Geometry*, 1930, pp. 110 *et seq.*, or D. M. Y. Sommerville, *Geometry of N Dimensions*, 1929, pp. 59 *et seq.*). The formula required above, however, may be more simply deduced as follows. Let x′ be the one-rowed vector $[x_1, x_2, \ldots]$, L the matrix for transforming x into x, and Λ the diagonal matrix of coefficients

$$\begin{bmatrix} \lambda_1 & 0 & \ldots \\ 0 & \lambda_2 & \ldots \\ \ldots & \ldots & \ldots \end{bmatrix}$$

Then, assuming it to be always possible to reduce $x'Mx$ to a sum of squares such as $\lambda_1 x_1^2 + \lambda_2 x_2^2 + \ldots$, we have

$$x'\,M\,x = x'\,\Lambda\,x$$
$$\text{or } x'\,L'\,M\,x = x'\,\Lambda\,L'\,x$$
$$\text{hence } M\,L = L\,\Lambda$$
$$\text{or } (M - \Lambda)\,L = O$$

which is the equation given in the text.

and will thus represent g's contribution to the total variance, i.e., $\Sigma(r_{kg}{}^2)$. For each value of λ we have the following set of equations to determine the proportionate values of the direction cosines :—

$$(M - \lambda I)\ L = 0, \text{ that is } (M_R - \frac{1}{\lambda} I)\, L = 0\ ;\ \text{or}$$

$$(r_{11} - \frac{1}{\lambda_1})\, l_{11} + r_{12}\, l_{12} + \ldots + r_{1n}\, l_{1n} = 0$$

$$r_{21}\, l_{11} + (r_{22} - \frac{1}{\lambda_1})\, l_{12} + \ldots + r_{2n}\, l_{1n} = 0$$

etc.

If the solutions of these equations be written $l_{11}P$, $l_{21}P$, ... , we can derive the saturation coefficients by the equation

$$r_{kg} = l_{11}\, P\, \frac{\sqrt{1/\lambda}}{\sqrt{\Sigma(l_{1i}{}^2 P^2)}}\ .$$

And similarly for the saturation coefficients for the more specific factors, r_{ks_i}, etc.

If we regard s_i, not as specific factors but as sources of error, the above formula for r_{kg} becomes identical with that given by equation (xxviii). Hence the formula which has been derived above on the assumption that all factors except the most general might be treated as negligible still holds good even if there are as many general factors (that is, common or partly common factors) as there are tests.[1]

582. We have now obtained a direct method for estimating how far the marking of any given examiner correlates with the hypothetical true marking at which presumably all are aiming. Taking the mark-lists for Latin by way of illustration, let us calculate the " h.g.f. coefficient " for each of the six examiners.

[1] This is in effect a simplified version of the second proof given by Hotelling, *loc. cit.*, pp. 426–7. If the argument as I have stated it above be set out in full for the cases of two variables and of three, it can easily be followed by the student with no knowledge of advanced mathematics ; and the extension to the n-dimensional problem becomes obvious.

Hotelling also offers an alternative form of his proof. Following the usual procedure for conditional maxima (Lagrange's method) he takes partial derivatives and introduces the λ's of the foregoing argument as undetermined multipliers and then evaluates them (*loc. cit.*, p. 423). It is instructive to note that this method can be readily applied to the construction based on T-axes with the s.d.'s for the tests set equal to unity. With this construction the requirement will be that the sum of the projections of all the tests on the g-axis shall be a maximum. The result is identical with that just reached by using T-axes. It differs from the simpler solution obtained for F-axes merely in using second moments instead of first.

The first step is to compute the intercorrelations between each examiner and the rest (Table 140).[1] The coefficients range from ·48 (a figure that is barely significant) to ·86, and average ·75. Had we a random sample of scripts instead of scripts selected on the basis of equal and approximately average marks, we should here doubtless have obtained somewhat higher coefficients.[2] Generally I find that, with unselected samples, the correlations range between ·6 and ·9 according to the nature of the subject marked. On experimenting with examinations for academic subjects by means of what are sometimes called "new type" question-papers (papers requiring a large number of short answers instead of a small number of long answers, and generally modelled on the lines of written tests for intelligence) I have obtained correlations between the marking of different examiners rising to nearly 1·00 and rarely falling below ·9.

[1] The averages of the observed correlations (Table 140, p. 293, last line but two) are, as we should expect, a little lower than the correlations with the average (Table 138, p. 277, top line). They thus furnish an alternative indication, though not a final indication, of the relative accuracy of the examiners as judged by a comparison of their marks with their colleagues'. These average correlations are, in fact, the figures which the French investigators use in calculating the relation between each examiner and *la vraie note*. (I am grateful to Sir Philip Hartog for drawing my attention to this point, and allowing me to see a copy of the advanced proofs of their report, since my own memorandum was written.) It appears that the French investigators have implicitly followed much the same line of reasoning as that set out above: indeed, they refer freely to the methods, and employ the formulæ, of Spearman and other English psychologists. Having calculated such tables as the above, instead of computing a "saturation coefficient" from the figures, they have used the simple average. They then employ the formula $\sigma_k\sqrt{1 - \bar{r}_{kk'}}$ to measure the standard error or deviation, *des mesures faillibles* (actual marks) *autour de la vraie valeur qu'elles visent:* (here $\bar{r}_{kk'}$ = the average correlation calculated as above). This formula takes the place of my formula $\sigma_k\sqrt{1 - r_{kg}^2}$. In my opinion it gives approximately the correct order of merit for the several examiners, but slightly overestimates the absolute amount of their "standard error."

The grand average (·75), it is interesting to note, is here of much the same size as the average found by the French investigators. Taking all the subjects for the Baccalauréat, 100 scripts for each subject, and 6 examiners for each script, they find that the average of the 165 correlations they have calculated amounts to ·71. It would be of great interest to compile similar averages for other subjects. In my own limited inquiries I find, for example, that where two examiners, an internal and an external, have marked the same scripts in complete independence, the correlations for papers on mathematics rise to ·91; those for English literature average only ·63; those for psychology ·58; and those for philosophy ·46. If we proposed to reduce the personal peculiarities of the markers by increasing their number, and took the accuracy of the mathematicians as our ideal, we should require, in a subject like philosophy, an internal board and an external board with at least 12 examiners on each before we could raise the correlation sufficiently to indicate the same degree of agreement between the internal marks and the external marks in that subject as obtains between two examiners in mathematics.

[2] See footnote 1, page 268 and § 568, *ad init.*

TABLE 140

MARKS FOR SCHOOL CERTIFICATE LATIN: GROUP I

ACTUAL AND THEORETICAL CORRELATIONS BETWEEN THE MARKS OF THE SEVERAL EXAMINERS[1]

H.g.f. coefficient determined by Summation Method

Exr.	H.g.f. coeff.	Examiner	F	A	B	E	D	C	Average
		H.g.f. coeff.	·951	·920	·910	·863	·855	·726	·871
F	·951	Actual		·860	·837	·815	·837	·708	·811
		Theoretical	(·904)	·875	·865	·823	·803	·686	·810
		Difference		−·015	−·028	−·008	+·034	+·022	+·001
A	·920	Actual	·860		·803	·742	·851	·705	·792
		Theoretical	·875	(·846)	·840	·798	·779	·665	·792
		Difference	−·015		−·037	−·056	+·072	+·040	·000
B	·910	Actual	·837	·803		·800	·812	·670	·785
		Theoretical	·865	·840	(·828)	·789	·770	·657	·785
		Difference	−·028	−·037		+·011	+·042	+·013	·000
E	·863	Actual	·815	·742	·800		·722	·688	·753
		Theoretical	·823	·798	·789	(·745)	·731	·625	·753
		Difference	−·008	−·056	+·011		−·009	+·063	·000
D	·855	Actual	·837	·851	·812	·722		·478	·740
		Theoretical	·803	·779	·770	·731	(·731)	·609	·739
		Difference	+·034	+·072	+·042	−·009		−·131	+·001
C	·726	Actual	·708	·705	·670	·688	·478		·650
		Theoretical	·686	·665	·657	·625	·609	(·527)	·649
		Difference	+·022	+·040	+·013	+·063	−·131		+·001
Average		Actual	·811	·792	·785	·753	·740	·650	·755
		Theoretical	·810	·792	·785	·753	·739	·649	·755
		Difference	+·001	·000	·000	·000	+·001	+·001	·000

Having calculated the intercorrelations, the next step will be to compute the h.g.f. coefficients by one of the methods described above (equations XXV, XXVI, or XXVIII). Here for the sake of comparison I have used all three—the product method, the summation method, and the method of least squares. The final

[1] For the observed or "actual" correlations, the probable errors, calculated by the usual formula, range from ± ·094 to ± ·032. But the latter slightly exaggerates the significance of the lowest coefficient. For $r = ·478$, $\frac{z}{\delta_z} = 1·96$. Assuming normal distribution, such a coefficient might arise about once in 20 times or just a little more frequently. Using the published tables for t (Fisher, *loc. cit.*, p. 196) and taking $P = ·05$ (1 in 20) as marking the borderline for significance, we find that with groups of this size the smallest significant coefficient would be ·514.

results are shown in Table 141. Since the original matrix of observed correlations is nearly hierarchical, all three determinations agree pretty closely. On averaging the squares of h.g.f. coefficients, we find that the "general factor" as thus determined accounts for ·764—rather more than three-quarters—of the total variance. The rest, therefore, must be attributable to more specific factors or to random errors.

TABLE 141

MARKS FOR SCHOOL CERTIFICATE, LATIN : GROUP I

CORRELATION BETWEEN THE MARKS OF EACH EXAMINER AND THE HYPOTHETICAL TRUE MARKS

Examiner	A	B	C	D	E	F	Average
H.G.F. Coefficient by							
Least Squares Formula	·920	·910	·726	·855	·863	·951	·871
Summation Method	·922	·911	·721	·844	·866	·950	·869
Product Method	·926	·914	·716	·831	·871	·955	·869
Probable Error	·028	·031	·085	·048	·045	·017	·043
Residual Error	·392	·415	·688	·519	·505	·309	·490

From any of these theoretical figures we can reconstruct the correlations to be expected between one examiner and another on the assumption that those correlations are due solely to the common influence of the true value of the scripts and in no way to special viewpoints shared by two examiners but not by all. Table 140 (p. 293) illustrates the method and gives the results. The theoretical figures in the body of the Table represent the matrix M_R obtained by multiplying M_1 and its transpose M_1'. Here M_1 and M_1' consist of but a single column and the corresponding row, which appear as the h.g.f. coefficients at the side and the head of the table, and may accordingly be briefly denoted by the symbols M_{kg} and M'_{kg} respectively. The theoretical correlations, therefore, are simply the products of these h.g.f. coefficients taken in pairs. In the table I have taken h.g.f. coefficients as obtained by the least squares method. The resulting correlations may be regarded as a kind of smoothing of those actually observed. Incidentally they serve to illustrate in the concrete what is meant by a hierarchical order.[1]

583. I have made similar reconstructions with each of the three formulæ for the h.g.f. coefficient. On subtracting these expected figures from the correlations actually observed, we can

[1] In the Table the examiners are rearranged in order of their average correlation to show how far the coefficients approximate to the "hierarchical" arrangement which, as we have seen, should result in the absence of overlapping group-factors. With probable errors so high, it is impossible to apply the tetrad-difference criterion.

discover which method yields the closest fit. The least squares and the summation method yield a total discrepancy of zero: the product method yields a total discrepancy of +·060—the reconstructed correlations being (as we might expect) somewhat enlarged by the process of continued multiplication. The total *square* deviation—perhaps the best criterion—is therefore also largest for the product method, namely, ·080; for the least squares method it is very slightly smaller than for the summation method—the ordinary Spearman saturation coefficient—barely ·074 as compared with ·075. With tables that approximate less closely to hierarchical order, the former generally shows a definite superiority.

584. In the main the new h.g.f. coefficients of Table 141 resemble the old correlations of Table 138: that is to say, the correlations with the true marks calculated by the more elaborate method do not greatly differ from those calculated by the rough and speedy method of correlating each examiner's marks with the unweighted average. Since an unweighted average already includes the mark we are correlating with it just as it stands, we shall not be surprised to find that the cruder method slightly magnifies the apparent correlation for those examiners (C, for example) who do not deserve so high a weight as the others. For practical purposes, however, when there is no great difference between the original correlations, the more rough and rapid method is sufficiently exact: it may certainly serve to determine the general accuracy of the examiners as a group, though not their individual differences.

Here, whether determined by the crude method of Table 138 or any of the more exact methods used for Table 141, the order of magnitude of the several coefficients remains practically the same. It differs from the order of efficiency obtained in the body of the Report (Part II, p. 187), in that the latter makes D the most unreliable examiner: that is simply because his wide and discriminating standard deviation has there been treated as part of his residual error or "random variation." In point of fact, however, if any distinction can fairly be drawn on the basis of so few candidates, C is the one examiner who markedly disagrees with the rest of his colleagues and so is presumably the least efficient.

Whether judged by his crude correlation with the average or by his h.g.f. coefficient, F appears the most efficient. But with the h.g.f. coefficients his superiority stands out more clearly. (In this respect, as elsewhere, the three different formulæ for the h.g.f. coefficients yield much the same result.) This leads

me to stress the fact that the judgment of the best judge necessarily correlates far more closely with the truth than would be suggested by his average correlation with other judges or his correlation with the average of them all.

Between F's coefficient and C's we now find a difference of ·225, considerably larger than that arrived at by the cruder method. This difference we may accept as just statistically significant[1]; but all the other differences are so small in comparison with the probable errors that they may be regarded as virtually negligible. With groups of about twice the size, however, containing say about 30 or 40 candidates, and with correlations of this order, it is evident that the differences would rise rapidly into prominence.

585. In the last line of Table 141 I give what I have called above (p. 278) the coefficients of non-relation. Here the coefficients, being based upon a more accurate estimate than the correlations with the average, show, rather more precisely than the figures then considered, to what extent each examiner has reduced the amount of error attributable to sheer chance.

TABLE 142

MARKS FOR SCHOOL CERTIFICATE, LATIN: GROUP I

RESIDUAL ERRORS IN TERMS OF ORIGINAL MARKS.[2] ("Standard Deviations of Random Variations")

Examiner	A	B	C	D	E	F	Average
Present Method	1·47	1·61	2·66	2·75	2·10	1·17	1·96
Approximate Method (para. 624, p. 321)	1·52	1·52	2·59	2·63	1·92	1·26	1·91
Approximate Method (para. 493, p. 225)	1·45	1·69	2·66	2·72	2·09	0·88	1·91

586. If we multiply these figures by each examiner's standard deviation, we shall obtain what is called in the body of the volume the standard deviation of his random variations—that is, the squares of the crude residual errors. The results are shown in Table 142. For comparison I append the corresponding figures as calculated by the approximate method described in Part II (para. 493, p. 225). It will be remembered

[1] Calculated by the ordinary formula directly from the coefficients, the probable error of the difference between C's correlation and F's is ± ·059. If instead we apply Fisher's criterion, we have $tanh^{-1}\cdot951 - tanh^{-1}\cdot726 = \cdot92$. The standard error of this difference is $\sqrt{\frac{2}{n-3}} = \cdot408$; and 3 times the probable error of the difference = ·835. The difference therefore is still significant.

[2] i.e., before transformation to terms of each examiner's standard deviation taken as unity.

that the estimates were there calculated by taking the unweighted average as a first approximation to the ideal; that procedure, as we have seen, tends to diminish the amount of disagreement between the less efficient examiners and the ideal. Making due allowances for the differences in method, and considering the small size of the group, the figures tally fairly well.[1] Where, however, the groups are large enough for the divergences to be significant, the more rigorous method should, in my view, be adopted.

So far as the order is concerned, that resulting from Dr. Rhodes' methods and my own is precisely the same. But it is not the order shown by the coefficients of correlation. This is because the large standard deviation of an examiner like D increases his apparent amount of inaccuracy. If D was wrong in spreading out his candidates in this way, then the figure in Table 142 correctly exhibits his resulting inaccuracy: if D's powers of discrimination, and his confidence in them, are justified, then the simple coefficient of non-relation as given in Table 141 (i.e., without multiplying by σ) affords the better measure. In any case, a full analysis should distinguish the two sources of the apparent error.

Section VI.—To Determine the Hypothetical "True" Mark for a Given Candidate: the Weighted Average

587. To estimate the true mark for each candidate we shall have to combine the marks awarded by each of the n examiners. If the examiners differ widely in their accuracy, as shown, for example, by their differing correlations with the general factor, then, instead of simply adding or averaging the several marks as they stand, we ought in theory to weight them first of all. What weights, then, are we to employ, and how much will our estimates gain in precision?

588. Let g' be the best estimate for the hypothetical "true" mark (g) for any given candidate, and $x_1, x_2, \ldots, x_n$ the marks actually awarded; then, instead of taking

$$g' = \frac{1}{n}(x_1 + x_2 + \ldots + x_n), \text{ we shall take}$$

$$g' = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n, \qquad \ldots \text{ (xxx)}$$

[1] The top line of figures in Table 142 is based on the h.g.f. coefficient as obtained by the least squares formula. At the corresponding point in his argument, Dr. Rhodes now uses what I have called the product method (equivalent to taking geometrical means): cf. pp. 285 and 317. This rather exaggerates the h.g.f. coefficient for F and accounts for the discrepancies in his figures for F in Table 142. For the reasons given above, I should be disinclined, except in special circumstances, to compare examiners upon this basis.

where the weights, w_k, are fractions chosen so as to keep (so far as possible) the standard deviation of g' equal to 1 and therefore equal to that of the marks of each examiner when given in standard measure.

To determine the weights, we may adopt the equation given at the outset (equation (ii), p. 247). This was based on the matrix M_R; and this, as we have seen, is simply M_R (which we may now legitimately identify with $M_{kk'}$) bordered by the h.g.f. coefficients. M_R as actually observed, however, may be imperfect, since the self-correlations may be missing or unduly augmented; but, if it is presumably produced by a single general factor, we may regard it as an empirical estimate for the theoretical hierarchy $M_{kg}.M'_{kg}$. In that case the determinants simplify; and it becomes possible to offer a more direct formula.

On this assumption, we have as before (cf. p. 270), for any given candidate, n equations of the form :—

$$x_{kj} = b_{kg} \cdot g_j + e_{kj} \qquad \ldots \text{ (xxxi)}$$
$$(k = 1, 2, \ldots n)$$

where e_k denotes the errors or random variations peculiar to each examiner. b_{kg} is known; we desire to estimate g_j. The simplest procedure is to follow the method of least squares as applied to the familiar problem of estimating a single quantity by combining observations of unequal precision. Square the expression given by the equation for the errors of estimate; differentiate in respect of g; equate the derivative to zero; then, substituting $r_{kg}\frac{\sigma_k}{\sigma_g}$ for b_{kg} and $\sigma_k^2(1-r_{kg}^2)$ for σ_e^2 (the standard deviation of the errors) we obtain

$$g = \frac{\Sigma \dfrac{r_{kg}}{1-r_{kg}^2} \cdot \dfrac{x_k}{\sigma_k}}{\Sigma \dfrac{r_{kg}^2}{(1-r_{kg}^2)\sigma_g^2}} \qquad \ldots \text{ (xxxii)}$$

The denominator is constant. Accordingly, it follows that to determine g, we must weight the marks of any given examiner k in proportion to

$$\frac{r_{kg}}{(1-r_{kg}^2)\sigma_k}$$

Thus, each weight varies with three factors, differing for the different examiners : (i) it is proportional to his h.g.f. coefficient—i.e., to the closeness with which he correlates with the true mark;

(ii) it is inversely proportional to the square of the probable error of his estimate; (iii) and it is inversely proportional to the standard deviation or general spread of his marks, if these have not already been converted into standard measure.

Or we may treat the weights as the unknowns, and differentiate in respect of them, following the method commonly used for a known criterion (see p. 247, above); the result is the same.

Or, without invoking the differential calculus, we may put $g' = g + e$ in equation (xxx); multiply by any set of marks, x_j say; sum for the N candidates and divide by N; put $r_{jk} = r_{jg}\, r_{kg}$; and then repeat the process, multiplying by a second set of marks, $x_{j'}$, say. On subtracting and collecting terms, we have

$$w_j : w_{j'} = \frac{r_{jg}}{1 - r_{jg}^2} : \frac{r_{j'g}}{1 - r_{j'g}^2} \quad \text{as before} \quad \ldots (\text{xxxiii})$$

589. Now, however accurately we weight the original marks, the result can never be more than an estimate of g, analogous to the estimate derived from a partial regression equation. Such estimates do not have the same standard deviation as the true values. Accordingly, if the true value is in standard measure, then, in order to obtain an estimate which shall also be in standard measure, we must discover the standard deviation of the estimated marks.

Let us employ simply the proportionate weights, calling them w'_k, and using g'' to designate the estimate so reached, i.e., $\Sigma(w'_k\, x_k)$. We at once obtain

$$\sigma_{g''}^2 = \{\Sigma(w'_k\, r_{kg})\}^2 + \Sigma w'^2_k(1 - r_{kg}^2)$$

Then, substituting from (xxxiii) and writing S for $\Sigma(w'_k\, r_{kg}) = \Sigma \frac{r_{kg}^2}{1 - r_{kg}^2}$, we have $\sigma_{g''} = \sqrt{S(S+1)}$; and, on dividing w'_k by this value, we at once reach the equation for w_k (xxxv).

Or, instead of using the proportionate weights, we may calculate the absolute weights directly. We then obtain

$$\text{estimated } g = r_{gg'} \cdot g' = \frac{1}{1+S} \Sigma \frac{x_k\, r_{kg}}{1 - r_{kg}^2} \quad \ldots (\text{xxxiv})$$

Again the estimates will not be in standard measure, but will have a standard deviation $= r_{gg'} = \sqrt{\frac{S}{1+S}}$ (as proved below; see equation (xxxviii)). This leads to the same result.

Thus, if we desire an estimate for g in standard measure, we must first reduce the original marks to standard measure and then take

$$w_k = \frac{r_{kg}}{1 - r_{kg}^2} \cdot \frac{1}{\sqrt{S(S+1)}} \qquad \ldots \text{(xxxv)}$$

590. It will be noted that the absolute weights implied by equation (xxxiv) are

$$\frac{1}{1+S} \cdot \frac{r_{kg}}{1 - r_{kg}^2} = W_k \text{ (say)} \qquad \ldots \text{(xxxvi)}$$

Now, if Δ is the determinant of M_R (as defined in para. 548) and if M_R (the matrix corresponding to Δ_{gg}) $= M_{kg} \cdot M'_{kg}$, then $r_{kk'}$ (as we have seen) $= r_{kg} \cdot r_{k'g}$. If further we assume that r_{kk}, the theoretical self-correlations in the leading diagonal, all $= 1$, then, on evaluating the determinants by Chiò's theorem in the usual way, most of the second order minors (or tetrad differences) vanish ; and, as a result of this simplification,

$$(-1)^{k+1} \frac{\Delta_{kg}}{\Delta_{gg}} = \frac{r_{kg}}{1 - r_{kg}^2} \cdot \frac{\Delta}{\Delta_{gg}}$$

$$\text{and } \frac{\Delta}{\Delta_{gg}} = \frac{1}{1+S}$$

$$\text{Hence } W_k = (-1)^{k+1} \frac{\Delta_{kg}}{\Delta_{gg}}$$

which is the value given by equation (ii), p. 247. Thus, under these conditions, the two formulæ are entirely consistent.[1]

591. The formula, however, thus deduced for W_k is derived by assuming the existence of but one common factor, namely, g. If now we assume the existence of many common factors instead of only one, it is still possible to generalize the deduction. The argument can be expressed most simply by putting it in matrix notation.

We are given M_3, the candidates' actual marks as awarded by the several examiners. We know $M_{kk'}$, the intercorrelations between the examiners' marks : this we have seen to be $= M_1 M'_1$, the product of the matrix of h.f. coefficients[2] with its transpose. We require to find M_2, the candidates' hypothetical

[1] As the conclusions given above show an apparent divergence from those reached by Dr. Rhodes, I have thought it best to indicate several alternative lines of proof. It may be added that for individual candidates no estimate can be infallible : it must always contain an indeterminate element, which, however, tends to vanish as the number of examiners is indefinitely increased.

[2] I use the phrase "h.f. coefficients" to denote the hypothetical weightings for the different elements or factors (see Table 134). Their matrix (M_1) only $= M_{kg}$ when there is but *one* element—the hypothetical "general" factor (g)—to be considered.

true marks for the several factors. In accordance with Table 134 (p. 250), we have

$$M_3 = M_1 M_2 \text{ and } M_1 M_1' = M_{kk'}$$
$$\text{therefore, } M_1 M_1' M_{kk}^{-1} = M_{kk'} M_{kk}^{-1'} = \mathrm{I}$$
$$\text{and } M'_1 M_{kk}^{-1'} M_3 = M_2$$

Hence the matrix of weights[1] required to estimate M_2 from M_3 is

$$M_W = M_1' M_{kk}^{-1'} \quad \ldots \text{ (xxxvii)}$$

It is easy to see that the argument based on the assumption of one common factor is but a special case of the argument just given: for in that case M_1 becomes the matrix of h.g.f. coefficients and so consists solely of the array $r_{1g}, r_{2g}, \ldots, r_{ng}$; and the equation for W_k indicated by (xxxvi) at once reduces[2] to the formula given in equation (ii).

592. Let us now use the formula we have deduced to obtain the weights appropriate to the various examiners in the examination for School Certificate Latin. Table 143 shows (1) the weights (w_k) to be applied to the marks (presumed to be *already* converted into standard measure) in order to obtain the best estimate (also in standard measure) and (2) the proportionate weighting $\left(\frac{w_k/\sigma_k}{\Sigma(w_k/\sigma_k)}\right)$ to be applied to the marks *before* conversion into standard measure: i.e., this latter figure takes into account not only the differences in the several examiners' h.g.f. coefficients, but also the differences in their original standard deviation.[3]

[1] cf. Hotelling, *loc. cit.*, p. 418, who only takes the simplest case where $q = n$. Equation (xxxvii) can also be deduced, perhaps a little more validly, by applying the method of least squares to the initial matrices. It would be instructive to relate M_2, as estimated by this equation, with the alternative M_2 directly obtainable by applying factor-analysis to the *columns* of M_3 (see p. 253). If M_3 is test-measurements, both describe the ultimate constitution of the persons tested, but the alternative M_2 depicts them as correlated, since the diagonal matrix of variances implicit in M_1 is now transferred to M_2.

[2] The reduction is obtained by remembering that M_{kk}^{-1}, the reciprocal or inverse of $M_{kk'}$, is the square matrix whose (k, k')th element is the cofactor of $r_{kk'}$ in the determinant $|r_{kk'}|$ divided by the determinant itself.

[3] The second line of the Table gives figures comparable to those given in the last line of Table 117 of the body of the Report (Part II, p. 187) for obtaining "a better approximation to the ideal." The chief difference is that F there has only 4 times the weight of C, instead of 7 times. With mark-lists such as the present the difference is wholly negligible. My object, however, is not to improve upon that table, but to illustrate a method which seems to me more expeditious than that of repeated and successive approximations by the repeated calculation of the "random" (i.e., residual) "variations."

It will be observed that F's marks merit nearly 7 times as heavy a weight as C's. To statisticians, however, it is a familiar fact that, even with large differences in weighting, a weighted average, as a rule, differs but little from the ordinary or unweighted average. Here, for example, when the two averages, weighted and unweighted, are placed side by side, the change that results from the weighting is so slight as to affect the decimals only. To save space I shall not print the mark-lists in detail, for a comparison by eye is of little value. We can, however, directly measure what little improvement there may be by a coefficient of correlation.

TABLE 143

SCHOOL CERTIFICATE, LATIN: GROUP I

THEORETICAL WEIGHTING TO BE APPLIED TO THE SEVERAL EXAMINERS' MARKS AFTER AND BEFORE CONVERSION INTO STANDARD MEASURE

Examiner	A	B	C	D	E	F
Weighting— after conversion (w_k)	·239	·199	·056	·109	·129	·363
before conversion $\left(\frac{w_k/\sigma_k}{\Sigma(w_k/\sigma_k)}\right)$	·231	·186	·052	·075	·112	·345

593. What, then, is the correlation of these estimated marks with the actual value of the true mark ? To answer this question we must determine $r_{g'g}$. For this purpose there is no need actually to compute the two mark-lists. We can obtain a general formula as follows :—

$$r_{g'g} = \frac{\overset{N}{\underset{1}{\Sigma}}(g'g)}{N\,\sigma_{g'}\,\sigma_g}\text{, or, substituting from (xxii),}$$

$$= \frac{\overset{N}{\underset{1}{\Sigma}}(w_1 x_1 \,.\, g) + \overset{N}{\underset{1}{\Sigma}}(w_2 x_2 \,.\, g) + \ldots + \overset{N}{\underset{1}{\Sigma}}(w_n x_n \,.\, g)}{N\,\sigma_{g'}\,\sigma_g}$$

$$= \overset{k=n}{\underset{k=1}{\Sigma}}(w_k\, r_{kg})\text{, since } \sigma_k, \sigma_g, \sigma_{g'} \text{ all} = 1.$$

It will be observed that the correlation of the estimates with the actual values thus proves to be the weighted sum of the several h.g.f. coefficients of each examiner. Continuing, and substituting the values for w_k given given by (xxxv), we obtain

$$r_{g'g} = \frac{1}{\sqrt{S(S+1)}} \Sigma \frac{r_{kg}^2}{1 - r_{kg}^2} = \frac{S}{\sqrt{S(S+1}} = \sqrt{\frac{S}{S+1}} \quad \ldots\ldots \text{(xxxviii)}$$

The formula gives for School Certificate Latin $r_{g'g} = \cdot 982$.

594. To obtain a corresponding correlation for the unweighted average we may use the familiar formula for correlation of sums :—

$$r_{\bar{k}g} = \frac{\Sigma \sigma_k r_{kg}}{\sqrt{\Sigma \sigma_k^2 + \Sigma \sigma_k \sigma_{k'} r_{kk'}}} \quad \ldots \text{(xxxix)}$$

where the summation of the correlations in the denominator is taken over the whole of the $n\ (n-1)$ coefficients and $\bar{k}$ stands for the unweighted average. For School Certificate Latin this formula gives $r_{\bar{k}g} = \cdot 975$.

If the standard deviations are approximately equal, the last formula reduces to

$$r_{\bar{k}g} = \frac{n \bar{r}_{kg}}{\sqrt{n + n(n-1)\bar{r}_{kk'}}} = \sqrt{\frac{n \bar{r}_{kk'}}{1 + (n-1)\bar{r}_{kk'}}} \quad \ldots \text{(xl)}$$

since, as we have seen, $\bar{r}_{kg} = \sqrt{\bar{r}_{kk'}}$ approximately. Hence we obtain a rapid method for estimating the probable correlation of a set of averaged marks with the true marks, obtained from n examiners whose average correlation with one another, $r_{kk'}$, is known. This approximate formula here gives the same figure as the full formula to the third decimal place, viz., $\cdot 975$.

595. Conversely, we may desire to know how many examiners would be wanted to obtain a set of marks that would possess a specified degree of accuracy. A simple formula for this purpose can at once be derived from the above : it yields

$$n = \frac{r_{\bar{k}g}^2 (1 - \bar{r}_{kk'^2})}{\bar{r}_{kk'^2} (1 - r_{kg}^2)} \quad \ldots \text{(xli)}$$

Suppose, for example, we wish to enlarge the board of examiners for School Certificate Latin so that the amount of error in a purely chance marking shall be reduced, not to one-half (the effect of a single examiner), but to one-tenth, i.e.,

$$r_{\bar{k}g} = \sqrt{1 - \frac{1^2}{10}} = \cdot 995$$

Then, if $r_{\bar{k}k} = \cdot 755$ (as here), we at once obtain $n = 74 \cdot 7$. Thus, on the simple principle that errors may be progressively neutralized by averaging results from different sources, a board of 75 examiners would be needed to obtain the degree of accuracy desired.

596. Let us now compare the effects of averaging with weights and without. The marks of the best examiner, F, correlate, as we have seen, with the ideal marking to the extent of ·950; the unweighted average of the six examiners correlate with it to the extent of ·975; and the weighted average to the extent of ·982.

The successive improvements are exceedingly small. A consideration of the formulæ and of the results thus reached suggests several practical conclusions. First, for ordinary purposes as distinct from theoretical inquiries, the process of weighting seldom yields an amendment in any way comparable with the labour involved. Except with large numbers or high correlations, the weights themselves can only be determined within a wide margin of error: hence their use is often more likely to impair than to improve the final result. Secondly, with boards of a reasonable size, the process of averaging often fails to improve to any large extent upon the marking already supplied by the most competent judge: it may at times even spoil it. Thirdly, the best results are obtained by combining the marks of two or more highly competent judges whose marking is independent of each other's: i.e., who have high h.g.f. coefficients, but low correlations with one another. The same holds true of subjects: the best results will be obtained by combining examinations or tests which correlate highly with the general ability to be measured but attack it from independent or divergent angles.

Section VII.—Specific Factors

597. In Table 140 there are one or two instances in which the theoretical coefficients show a noticeable divergence from the observed coefficients. The largest divergences are those for the correlations of D with A and C respectively. Were these divergences significant we might be tempted to infer that the special standpoints of D and C were in some ways antagonistic, and that D and A perhaps possessed views in common which their colleagues did not share. Here, with so small a group, the high probable error forbids us to attach any real significance to these particular divergences. But with larger groups and

with different sets of figures, the application of one or other of the criteria suggested above (p. 282) might render it wholly impossible to account for the original matrix of observed correlations by assuming one general factor only. Other factors more specific, common to two or more examiners though not perhaps to all, might have to be postulated; and it will be of interest to inquire how these can be determined and measured.

598. The maximum number of factors is given by the number of columns in the weighting matrix (M_1). Hence, in theory, as we have seen, the minimum number of factors required to explain a given set of correlations may be deduced by examining the rank of the set considered as a matrix ($M_{kk'}$). In practice it will rarely be feasible to determine more than two or three factors.

599. The first step is to eliminate the influence of the general factor. The necessary equations are given by the result of multiplying the generating matrices (M_1 and M_1'): or, more simply, we may argue as follows.

The single factor theory assumed $r_{1g} = \sigma_g / \sqrt{\sigma_g^2 + \sigma_{e_1}^2}$, with a similar equation for Examiner 2. If we introduce a specific factor, s, we must alter the expression for the total variance which appears in the denominator, and write $r_{1g} = \sigma_g / \sqrt{\sigma_g^2 + \sigma_s^2 + \sigma_{e_1}^2}$. But now we can no longer add a similar expression for Examiner 2, since we cannot assume that the proportions of σ_g^2 and σ_s^2 are the same for both examiners. We must insert multipliers indicating these relative proportions, and take $r_{1g} = G_1 \sigma_g / \sqrt{G_1 \sigma_g^2 + S_1 \sigma_s^2 + E_1 \sigma_{e_1}^2}$ (say). But if these multipliers are chosen so that the several component variances σ_g^2, σ_s^2, σ_e^2, and their weighted sum, all remain = 1, we have $G_1 = r_{1g}$, $G_2 = r_{2g}$, $S_1 = r_{1s}$, $S_2 = r_{2s}$, where $r_{ks} \equiv$ the "saturation coefficients" for the specific factor s.

$$\text{Hence } r_{12} = \frac{\Sigma(x_1 x_2)}{N}$$
$$= (G_1 g + S_1 s + E_1 e_1)(G_2 g + S_2 s + E_2 e_2)$$
$$= G_1 G_2 + S_1 S_2 = r_{1g} r_{2g} + r_{1s} r_{2s} \quad \ldots \text{(xlii)}$$

Accordingly, to eliminate the influence of the general factor, we subtract the theoretical intercorrelation from the observed (as in finding the numerator of a partial correlation): i.e., we calculate the residuals

$$r_{kk'} - r_{kg} r_{k'g} \quad \ldots \text{(xliii)}$$

600. The next step is to determine whether these differences are statistically significant as judged by the probable error.

When the significance of one or more is proved, two procedures are possible. If we desire to estimate what specific correlation would obtain between two examiners in a population entirely homogeneous as regards the true mark, we shall divide the residual correlation by the residual variances. This is equivalent to applying the familiar formula for partial correlation.[1] If, on the other hand, we desire to measure how far the remaining factors are contributing to the total variance as actually observed, we shall take the residual correlations as they stand, working with the simple difference given by equation (xliii).

In either case the coefficients thus obtained from the whole Table will, like deviations from an average, show nearly equal totals of plus and minus values, and add up to a grand total of 0. In actual practice, they generally include a few high positive correlations and a large number of low and possibly insignificant negative correlations.[2] Accordingly, we may now look either (a) for specific factors whose influence is solely positive (the natural procedure in analysing mental abilities) or (b) for bipolar factors, which may account, not only for special resemblances between the examiners, but also for special antagonisms. For (a) the natural procedure is to take the positive specific correlations in sub-groups, and form separate matrices. The h.s.f. coefficient (saturation coefficient for specific factors) can then be calculated for each sub-group. For (b) we may convert the minus signs into pluses throughout a given row or column on the principle formulated above; namely, that if we are looking solely for influences as such, it does not matter whether the influence is negative or positive. By this means it may be possible to reduce still further the number of significant negative correlations. We can then seek as the second factor the component which is now contributing most to the total residual variance. This will usually be a bipolar "universal" factor, i.e., one apparently affecting all examiners, but some negatively and others positively. The process can be repeated until all the original variance is accounted for, or the residuals are statistically insignificant.

[1] $$r_{kk'.g} = \frac{r_{kk'} - r_{kg}\, r_{k'g}}{\sqrt{(1 - r_{kg}^2)\,(1 - r_{k'g}^2)}} \qquad \ldots \text{(xliv)}$$

which assumes, t will be observed, $\Sigma r_{ks} = \sigma_s/\sqrt{\sigma_s^2 + \sigma_e^2}$, i.e., omits σ_g^2 from the total variance: cf. p. 274.

[2] cf. *Distribution and Relations of Educational Abilities*, pp. 56 *et seq.* If with Thurstone we continue to apply equation (xxv), we definitely assume $\Sigma r_{ks} = 0$ exactly, with a consequent simplification (perhaps an over-simplification) of the resulting regression equations, as noted on p. 287 (footnote 1).

601. If reliability coefficients are obtainable and are inserted in the matrix $M_{kk'}$ before proceeding to calculate specific factors, the result will generally be that the first specific is a factor which is almost individual and pre-eminently characteristic of the examiner (n say) who had the lowest saturation coefficient for the general factor; it will show a high and positive specific saturation coefficient in his case, but diminishing and negative coefficients for $n - 1$, $n - 2$, ... , 2, 1 (that being their inverse order for the h.g.f. coefficients for the general factor). The next specific will be almost an individual factor characteristic of the $(n - 1)$th examiner; it will yield zero coefficients for n and diminishing negative coefficients for $n - 2$, $n - 3$, ... , 2, 1. If all the reliability coefficients are high and approximately equal, the process will continue; and the number of zero coefficients and the size of the negative coefficients will increase, while the number and size of the positive coefficients diminishes. The total result will be a triangular matrix of factor coefficients resembling the Jacobian canonical form. Here the unusual arrangement is a consequence of two peculiarities that characterize the procedure: first, the figures taken for the self-correlations are higher than non-individual factors alone would warrant, and secondly, all and perhaps more than all that can legitimately be regarded as common to all the examiners is absorbed at the outset into the first general factor.

It follows that, when mark-lists are analysed in this way, the specific factors peculiar to small groups of examiners, like the individual factors peculiar to one alone, have the effect of interference-factors—preventing the examiners who are influenced by them from agreeing with the true or general marking. Such a result would be difficult to accept were we dealing with mental abilities; but it is quite intelligible as an account of the influences affecting examiners' marks.

A similar triangular matrix results from a second method that can be used to analyse the correlation table into as many factors as there are tests.[1] Assuming that the correlations are arranged in order of their averages, the first examiner who heads the list may be regarded as giving a near approach to the general factor. Accordingly, taking his reliability coefficient r_{11} to be equal to r_{1g}^2, r_{1g} can be tentatively computed; and then r_{12}, r_{13}, ... may be divided by r_{1g} as thus determined in order to find r_{2g}, r_{3g}, ... The reliability coefficient of examiner 2 can be treated as equal to $r_{2g}^2 + r_{2s}^2$; and r_{2s} thus found from the residual $r_{22} - r_{2g}^2$. Dividing

[1] The procedure is based on a theorem given by Camp, *Biometrika*, XXIV, 1932, p. 422.

the other residuals by r_{2s} gives r_{3s}, r_{4s}, ... Then r_{ss} can be treated as equal to $r_{3g}^2 + r_{3s}^2 + r_{3ss}^2$; and the process continued. It is instructive to compare the results of the two procedures on an ideal hierarchy of coefficients with ideal coefficients (1·00) inserted for the reliabilities. Such calculations illustrate how artificial the analysis of specific factors is apt to be unless assumptions based on non-statistical or external considerations are made about the factors to be sought; and it is obvious that the specific factors rest on an entirely different footing from the general.[1]

602. It need hardly be added that for any analysis of specific factors to be worth the labour it is essential that the observed correlations should have high reliability coefficients and low probable errors—i.e., be obtained from examiners who are highly consistent with themselves and from large numbers of candidates.

[1] In this respect I agree with Spearman rather than with Hotelling or Thurstone. It is true we may for convenience of calculation treat the specific factors as simply additional general or universal factors; but their character and even their method of calculation is much more open to question. The difficulties are far more serious in the case of mental testing than of examining. In the latter we are primarily concerned to disentangle the temporary influences affecting one particular set of marks. In the former we are seeking permanent mental factors which shall be constant from one set of data to another. Even so to some extent the general factor is bound to be relative to the set of tests we choose. Obviously if we choose emotional instead of cognitive tests, we shall get different general factors; and the discovery that in dealing with different sets of cognitive tests the general factor varies so little is an empirical fact of psychology, not a result of the mathematical method, which can never go outside its data. The specific factors are still more dependent on the particular combination of tests chosen. If, in order to obtain a better fit for the general factor, we give a slight rotation to our ascertained axes, the distribution of specific factors may be greatly altered: and the addition or change of a single test may produce a very different set of specifics.

There seem to be two alternatives. First, we may seek a general factor which gives not the average slope for the plane of correlations but the basic slope: that is, we may seek to reconstruct a hierarchy in which no coefficient shall be lower than the corresponding observed coefficient (or the coefficient plus 3 times its *p.e.*) This would avoid specific factors with negative h.f. coefficients. Secondly, instead of changing our plan of analysis we may change our plan of testing. Instead of devising methods to discover as many hypothetical factors as possible from one set of test data, we may arrange the test data in the hope that they will be explicable by as few factors as possible—preferably only two. Thus we shall group our tests into two sets—one giving reference-values for the general factor, the other giving data for one additional specific factor only. Reliability coefficients, which will inevitably add as many factors as tests must be dropped from the analysis. The matrices involved will be a little anomalous; but it is not difficult to reduce them to general principles. Such methods will become relevant to work upon examinations as we pass from investigating the validity of given sets of marks to investigating either the personal characteristics of examiners as individuals or the inter-relations of examination subjects.

Section VIII.—Summary

603. The main conclusions to be drawn from the foregoing discussion are the following :—

(1) *Quantitative methods* of investigation are already available for studying the accuracy of marking in examinations of a scholastic or academic type : these methods will indicate how closely the marks of a given examiner or board of examiners are approximating to the hypothetical true marks and how far each is influenced by irrelevant factors of various types.

(2) Under certain specifiable assumptions, which can be approximately verified in actual practice, the methods of *factor analysis*, worked out for researches upon the validity of mental and scholastic tests, may, with slight modifications, be applied to the investigation of examination results. The simplest assumption is that the marks of any given examiner may be resolved into two hypothetical components—(i) the true value of the work to be marked, a component influencing all examiners but in different degrees, and (ii) residual errors, a component independent of the first and peculiar to each examiner. The difficulties that embarrass the investigation of mental factors for the most part do not arise in investigating examination results. In particular, by adopting a broader mathematical basis (treating the variables as co-ordinates in hyperspace) it can be shown that the seemingly divergent formulæ hitherto put forward are in their essential nature merely variants or alternative simplifications of one general conception. (Cf. paras. 552-556, 558-562, 566, 574-581).

(3) It appears that the simpler and speedier methods give *reasonable approximations* to a true result. With the rough data at present furnished by examination marks the more elaborate methods would be out of place. Such methods are in the main of theoretical interest, enabling us to justify, and occasionally to correct, the results secured by the more rapid short cuts. (Cf. paras. 583-586 and 592).

(4) The degree to which an examiner's marks agree with those of his colleague or colleagues may readily be measured by a *coefficient of correlation*, which can be calculated (if a simple and approximate formula be used) in a few minutes (equation (ix a), p. 271). (Cf. paras. 566-573).

(5) The discrepancies between the marks awarded for the same scripts by different examiners may be due, not only to imperfections in their powers or modes of judgment, but also to differences of scale, i.e., to individual differences in the general

standard of severity and in the degree to which the marks are spaced out; these differences can be measured, and if necessary allowed for, by calculating *averages and standard deviations.* (Cf. paras. 563-565).

(6) In theory the accuracy of a given examiner may best be measured by calculating a *general factor coefficient*, that is, by estimating the degree to which his marking correlates with the hypothetical true value taken as a standard. When the intercorrelations have been calculated, such coefficients can be determined at once by means of a simple formula (equation (xxv), p. 284). (Cf. paras. 574-586).

(7) Where the correlations between the several examiners and the true marks are not likely to differ widely, the *unweighted average* of the marks allotted by all the examiners yields a fair and quicker estimate of the ideal or true mark. (Cf. paras. 587-596).

(8) From the examiner's correlation with the true mark may at once be derived a *coefficient of non-relation* which measures the relative amount of random variation characterizing his mark: i.e., the degree to which he has failed in eliminating the influence of sheer chance. (Cf. paras. 571 and 585).

(9) Where the correlations between the several examiners and the true marks differ widely, *the marking of the best examiner* is almost as accurate as the average marking of the whole Board, and may even be more accurate. Accordingly, in certain types of examination, it may prove easier to increase the accuracy of the marking by trying to increase the accuracy of one or two examiners than by increasing their number and then averaging the results. Multiplying the number of examiners is of greatest value when their correlation with the true mark is high and their correlation with each other is low. (Cf. paras. 593-596).

(10) With small groups the calculations have a large *probable error.* But this can readily be determined; in such cases the newer statistical methods appropriate to small groups should be used. Such "errors," though high, do not necessarily invalidate the inferences so drawn; but they indicate the degree of significance which can safely be attached to them. (Cf. paras. 564, 570, 572, 582 and 584).

Finally, I would venture to urge that every examiner should acquire, as part of his training, some *knowledge of the elementary statistical principles* involved in such work. Every

examining body and every chairman of an examining board knows how each new member, fresh to his task, has slowly to pick up the requisite technique. There is a growing custom for preliminary instructions to be defined and set down on paper; and it would be a great advantage if the experience of expert examiners generally could be formulated, systematized, and recorded for the benefit of the novice. To be of real value, however, such instructions must, so far as possible, be couched in exact and quantitative terms. The art of examining, as of all forms of mental assessment and measurement, rests on scientific principles and involves a scientific technique; and this technique must, to a large extent, be statistical. Even so simple a task as combining marks for different questions cannot rightly be undertaken without a knowledge of what must occur when marks, more or less uncorrelated, are averaged or summed. Nor can a proper allocation of marks be made without some explicit theory as to the curve of distribution[1] which the marks should approximately follow. Again, examining authorities continually claim that the innovations which they have introduced must to a large extent eliminate the errors and uncertainties to which examinations in the past have been liable; but rarely if ever do they attempt a statistical experiment to verify how far the expected improvement has, in fact, been achieved. In the preceding pages I have limited my discussion solely to problems of consistency—i.e., to the correlations of examiners amongst themselves. This must be the first step in all such inquiries. But it is equally essential to investigate the accuracy of the results as judged by some external criterion—i.e., the correlation of the marks with independent first-hand evidence as to the merits of the candidates in the particular field with which the examination is concerned. These two lines of inquiry are always followed in estimating the value of any psychological test; and the same procedure should be adopted for every examination. Hence, in the present condition of our knowledge,

[1] For question-papers where the marking must largely turn on the subjective impressions of the examiner—those requiring answers of an essay type, for example—there is no device which increases accuracy and comparability so effectively as the plan of defining marks and their relative frequencies in terms of a distribution curve. But when any suggestion to this effect is made, almost invariably the normal curve is put forward as the ideal; and critics therefore rightly object that, since most higher examinations are taken by a selected batch of candidates only, a normal curve must in theory be inapplicable. The proper reply is that, given adequate data, it is always possible to find a better fitting curve—e.g., by taking some form of hypergeometrical curve (of which the normal curve is only a limited case) or by adjusting the normal curve itself by means of a logarithmic scale. In practice, however, it is usually found that selection (no doubt because it is so imperfect) introduces very little asymmetry or distortion.

the one thing needful is further research: and it is hoped that the foregoing review of the statistical principles involved and the statistical methods available may, in some small measure, be of assistance to this end.[1]

NOTE TO MEMORANDUM I

604. I have to acknowledge the kindness of several members of the Committee, in particular that of Sir Philip Hartog, Professor Hamley and Professor Godfrey Thomson,[2] who have been good enough to read my page proofs and to suggest several additions and corrections.

605. With Dr. Rhodes' memorandum, which follows, I have been unable to deal in the text. It would seem that the differences between us are now very slight.

In an early draft for the Committee I ventured to put forward equation (viii) (para. 566, p. 271 above) as a slight refinement on Dr. Rhodes' initial equation (para. 401, p. 190). My ground for this suggestion was that it seemed scarcely justifiable to assume that the standard deviation of each examiner's marks must always be the same. If their standard deviations are not equal (or approximately so within the limits set by the probable error), an important corollary would ensue: by an obvious algebraic proof, it was shown that we can no longer simply set the examiner's actual mark equal to the "true" mark plus an

[1] I have not thought it necessary to set out references or acknowledgments in full. It will be obvious that I am throughout deeply indebted not only to the writings of Professor Spearman (especially to his early article on "General Intelligence Objectively Measured," in the *American Journal of Psychology*, 1904, pp. 202–292, from which almost all this work has grown, and to his later book on *The Abilities of Man*), but also to the important series of articles in the *British Journal of Psychology* by Drs. Maxwell Garnett, Godfrey Thomson, Irwin, Piaggio, and others, to which the inquiring reader may profitably refer. The two most important American contributions (Hotelling, *Journ. Educ. Psych.*, XXIV, 1933, pp. 417–441, 498–520; Thurstone, *Psych. Rev.*, XXXVIII, 1931, pp. 406–427, and XVI, 1934, pp. 1–32) I have already cited. To British readers Thurstone's latest method is most accessible in Alexander's monograph quoted above. Thurstone's *Vectors of the Mind* (Univ. Chicago Press, 1935) unfortunately appeared only as these pages were being revised for the press. It contains a valuable and systematic summary of Thurstone's own treatment of the problem, which is in its essence a generalization of Spearman's, but would appear somewhat to magnify the divergences between Thurstone's methods and those of other investigators.

[2] Professor Spearman was unfortunately out of the country during the summer of 1935. With his consent, however, Dr. W. Stephenson, his former assistant, was good enough to read and check on his behalf the references to his work.

error; we must (even on the simplest hypothesis) set it equal to a *proportion* of the " true " mark plus a residual error. This means introducing a coefficient or multiplier, as is done in equations (vii), (viii) and (xxxi). Treating the actual mark as a *weighted* sum of two components led at once to a " two-factor " method of analysis, analogous to that derived by Professor Spearman from what he has termed the " two-factor theory " in the interpretation of psychological tests. But in my view even this slightly more elaborate equation is valid only under the simplifying assumptions made in the text.

Dr. Rhodes' original equation and my own therefore differ merely by my insertion of the coefficient r_{kg}. Dr. Rhodes, I gather, would now accept the introduction of such a coefficient into his formula (see Note III, p. 198). But, as he has shown in his memorandum (Appendix II, para. 629), with data like the present, the results obtained by his " new method " do not diverge significantly from those obtained by his old. With this conclusion I agree *where small groups only are available.*

In order to compare Dr. Rhodes' results with my own, I have inserted in the second line of Table 142 the figures that he has since calculated by the " new method " (Appendix II, para. 624 : as he applies it, it differs from mine chiefly in the adoption of the " product method " of calculating the h.g.f. coefficient instead of the " summation "). It will be seen that, with the mark-list chosen for illustration, the results obtained by all calculations show practically no significant differences.

606. I may add that this method of analysis is applicable to problems encountered in other sciences besides psychology—to problems of biology, economics, industry, and the like, where it is continually necessary to pool the figures given by different methods of assessment or to estimate the relative reliability of each method, and even to problems in the more exact science of physics. Conversely, psychologists might find much that is fruitful in the new mathematical methods used, for example, by contemporary physicists.[1]

607. Finally, a brief reference must be made to an important contribution by T. L. Kelley which appears as these pages are

[1] For such problems as we have been examining here, the tensor calculus (as developed, for example, in the study of relativity) and the theory of linear operators (as applied in quantum theory) would probably furnish most useful tools. (For the application of tensor calculus to statistical problems, see a suggestive paper by Eddington, *Proc. London Math. Soc.*, ser. ii, vol. xx, 1921, pp. 213-221 ; for the use of matrices and linear operators in problems of probability, see Weyl, *Theory of Groups and Quantum Mechanics*, 1931, and Neumann, *Mathematische Grundlagen der Quantenmechanik*, 1932.)

being passed for press.[1] His arguments proceed from much the same assumptions and follow much the same principles as those set out above. Although his method has been independently derived, the outcome of his " new technique " is, as he himself points out, " identical with that given by Hotelling ". On recalculating by means of Kelley's technique the coefficients for the sample examination taken above, I obtain figures which are practically identical with those given by the " least squares formula " (p. 294).

[1] *Essential Traits of Mental Life: The Purposes and Principles underlying the Selection and Measurement of Mental Factors*, Harvard University Press, 1935. The chief differences between Kelley's arguments and those of the foregoing memorandum appear to be as follows: first, he prefers to work with variances and covariances (a procedure certainly in keeping with recent trends in modern statistics) rather than in terms of standard deviations and coefficients of correlation—terms that still remain more familiar to educational psychologists: secondly, and partly it would seem as a result of this departure, he finds striking divergences between the results of his own method and those of Thurstone as he applies it: he concludes, a little too sweepingly in my opinion, that Thurstone's method must be regarded not merely as another and possibly a rougher method of approximation, but as resting on " logical foundations that are irreconcilably different "; lastly, since his " first component " (my " hypothetical general factor ") cannot be determined until all the other factors have been determined by successive approximations, his method, as a practical technique, proves decidedly slow and laborious for problems such as those we have been considering here.

MEMORANDUM II

A SECOND APPROXIMATION FOR THE DETERMINATION OF IDEAL MARKS AND RANDOM VARIATIONS

By
E. C. Rhodes

608. In Part II, para. 401, we have assumed that a mark X awarded to the tth piece of work by an examiner A is equivalent to $Q_t + A_t$, and that the mark Y_t awarded to the same piece of work by an examiner B is equivalent to $Q_t + B_t$, and so on. Q_t was designated the ideal mark, and A_t, B_t, etc., indicated by how much the various examiners diverged from the ideal. A more refined assumption would be

$$X_t = r_a Q_t + A_t,$$
$$Y_t = r_b Q_t + B_t \text{ and so on.}$$

On this assumption we should allow for the possibility that the various examiners might have different notions of the " spread " of the ideal marks, the multipliers r_a, r_b, etc., being different for the various examiners. This new assumption would include the original assumption as a special case when the multipliers r_a, r_b, ... are all equal.

609. In those cases where examiners are accustomed to the work of such examinations as the Special Place Examination, the School Certificate Examination, and Degree Examinations, where team work is a feature of the examinations and where instructions to examiners include such details as the limiting marks for the various grades, Failure, Pass, Credit, Second Class, First Class, etc., besides other detailed instructions as to the method of marking, we might reasonably anticipate that the examiners would agree on the ideal mark to be awarded to a script and therefore that this difference of " spread " suggested by the multipliers r_a, r_b, ... would not exist.

610. On the other hand, it is conceivable that when examiners are not working to instructions such differences as these might occur. At any rate it was thought worth while to test the new

hypothesis on some of the material of the investigation and to compare the results obtained with those given in Part II.

611. Starting then with the assumption

$$X_t = r_a Q_t + A_t,$$

we have to develop a new technique in order to estimate r_a, Q_t and A_t from our knowledge of the original marks. Proceeding as before and using the same notation, we have

$$x_t = r_a q_t + a_t$$

where x_t, q_t, a_t are deviations from averages.

Similarly $$y_t = r_b q_t + b_t.$$

From these we have

$$x_t^2 = r_a^2\ q_t^2 + 2\ r_a\ q_t\ a_t + a_t^2,$$

$$\text{and } x_t\ y_t = r_a\ r_b\ q_t^2 + r_a\ q_t\ b_t + r_b\ q_t\ a_t + a_t\ b_t.$$

If we sum all such expressions as these for values of t from 1 to n we have[1]

$$S(x_t^2) = r_a^2\ S(q_t^2) + S(a_t^2)$$

$$\text{and } S(x_t\ y_t) = r_a\ r_b\ S(q_t^2).$$

assuming that $S(q_t\ a_t) = 0$, $S(q_t\ b_t) = 0$, $S(a_t\ b_t) = 0$.

612. We can obtain estimates of r_a, r_b, etc., by using equations of the type $S(x_t\ y_t) = r_a\ r_b\ S(q_t^2)$.

If we repeat the process we get $\frac{m(m-1)}{2}$ equations of this type when there are m examiners. There are m quantities r_a, r_b, ... which are the *desiderata*. Normally these $\frac{m(m-1)}{2}$ equations will not be satisfied exactly by values of r_a, r_b, ... calculated from m of them. This is because the quantities $S(q_t\ a_t)$, $S(q_t\ b_t)$, ... , $S(a_t\ b_t)$, $S(a_t\ c_t)$, ... are not exactly zero. Our problem is to obtain estimates of the r's which will satisfy the $\frac{m(m-1)}{2}$ equations approximately. When we have obtained these estimates, which we do not pretend are the *best* estimates by any judgment of " best " which could be adduced, we propose

[1] We also have $x_t\ q_t = r_a\ q_t^2 + q_t\ a_t$.

Summing again, we have $S\ (q_t\ x_t) = r_a\ S\ (q_t^2)$.

If the correlation between X and Q is indicated by the correlation coefficient r_{qx},

$$r_{qx} = \frac{S\ (q_t\ x_t)}{\sqrt{S\ (q_t^2)\ S\ (x_t^2)}} = r_a \sqrt{\frac{S(q_t^2)}{S\ (x_t^2)}}$$

So our multiplier r_a may be considered as proportional to the correlation coefficient between X and Q.

to obtain estimates of $S(a_t^2)$ *et al.*, using the approximate equation $S(x_t^2) = r_a^2\, S(q_t^2) + S(a_t^2)$ and similar equations. With these approximations we shall be enabled now to obtain estimates of the ideal marks of each examiner. From these we can deduce better approximations to the $S(a_t^2)$ *et al.* Thus by successive approximations we can get reasonably correct estimates of the unknown quantities whose value we desire to establish.

613. A simple method of using the approximate relationships which we have established, for our purpose, is to multiply all the relationships

$$\begin{aligned} r_a\, r_b\, S(q_t^2) &= S(x_t\, y_t) \\ r_a\, r_c\, S(q_t^2) &= S(x_t\, z_t) \\ \vdots \quad \vdots &\qquad \vdots \quad \vdots \\ \vdots \quad \vdots &\qquad \vdots \quad \vdots \end{aligned}$$

which contain r_a on the left-hand side. There are $(m-1)$ such, and we get

$$r_a{}^{m-1} r_b\, r_c \ldots (S(q_t^2))^{m-1} = P_1 \text{ (say)}$$

where $P_1 = S(x_t\, y_t).\ S(x_t\, z_t) \ldots$.

We shall have similarly

$$r_b{}^{m-1}\ \ r_a\, r_c \ldots (S(q_t^2))^{m-1} = P_2 \text{ (say)},$$

where $P_2 = S(y_t\, x_t),\ S(y_t\, z_t) \ldots$.

Now, if we multiply all the $\frac{m(m-1)}{2}$ expressions

$r_a\, r_b\, S(q_t^2)$, we shall have

$$(r_a\, r_b\, r_c \ldots)^{m-1}\ (S(q_t^2))^{\frac{m(m-1)}{2}} = S(x_t\, y_t)\, S(x_t\, z_t) \ldots S(y_t\, z_t) \ldots = P \text{ (say)}.$$

Then $r_a\, r_b\, r_c \ldots (S(q_t^2))^{\frac{m}{2}} = P^{\frac{1}{m-1}}$

Thus $r_a{}^{m-2}\ \ S(q_t^2)^{\frac{m-2}{2}} = P_1/P^{\frac{1}{m-1}}$.

and we get $r_a \sqrt{S(q_t^2)} = P_1{}^{\frac{1}{m-2}} / P^{\frac{1}{(m-1)(m-2)}}$

Similarly, $r_b \sqrt{S(q_t^2)} = P_2{}^{\frac{1}{m-2}} / P^{\frac{1}{(m-1)(m-2)}}$ etc.

614. These are the first approximations from which estimates of r_a, r_b, ... are obtained. It will be noted that the r's are not given as absolute values. In fact it is impossible to divorce

the r's from $\sqrt{S(q_t^2)}$, and for convenience in the further work it is preferable to designate

$$r_a \sqrt{S(q_t^2)} \text{ by } R_a,$$
$$r_b \sqrt{S(q_t^2)} \text{ by } R_b, \text{ etc.}$$

615. The next stage is to obtain approximations of $S(a_t^2)$, etc. We use the approximate equations

$$S(a_t^2) + r_a^2 S(q_t^2) = S(x_t^2), \text{ etc.}$$

From them we get $S(a_t^2) = S(x_t^2) - R_a^2$.
$S(b_t^2) = S(y_t^2) - R_b^2$.
. . . .
. . . .

These give our approximations to $S(a_t^2)$, $S(b_t^2)$, ... since the right-hand sides of these equations are all now known.

616. Now we wish to use the approximate values thus obtained in order to estimate the ideal marks of each examiner. Let us consider the case of the marking of a single piece of work, the tth. Examiner A, when assessing this, introduces into his assessment an error a_t, of a system of errors with a standard deviation s_a. Examiner B introduces an error b_t, of a system of errors of standard deviation s_b, and so on. The chance of obtaining the errors introduced by the examiners is proportional to

$$e^{-\frac{1}{2}\left(\frac{a_t^2}{s_a^2} + \frac{b_t^2}{s_b^2} + \ldots\right)}$$

assuming that these error systems are normally distributed.

This expression is equal to

$$e^{-\frac{1}{2}\left(\frac{(x_t - r_a q_t)^2}{s_a^2} + \frac{(y_t - r_b q_t)^2}{s_b^2} + \ldots\right)}.$$

In order to find the most likely q_t to associate with this particular script we find the value of q_t which will make the chance referred to above a maximum. Thus we desire to know what value of q_t will make

$$\frac{(x_t - r_a q_t)^2}{s_a^2} + \frac{(y_t - r_b q_t)^2}{s_b^2} + \ldots \text{ a minimum.}$$

This value of q_t is given by differentiation, and is

$$q_t = \frac{x_t \frac{r_a}{s_a^2} + y_t \frac{r_b}{s_b^2} + \ldots.}{\frac{r_a^2}{s_a^2} + \frac{r_b^2}{s_b^2} + \ldots.}$$

It will be observed that when r_a, r_b, ... are all equal to unity, this expression reduces to the form used in Part II. Actually since we do not know r_a, r_b, ... but know R_a, R_b, ... it is preferable to write the above, substituting for s_a^2, etc., $S(a_t^2)$, etc., since $s_a^2 = S(a_t^2)/n$, in the form[1]

$$\frac{q_t}{\sqrt{S(q_t^2)}} = \frac{x_t \dfrac{R_a}{S(a_t^2)} + y_t \dfrac{R_b}{S(b_t^2)} + \cdots \cdots}{\dfrac{R_a^2}{S(a_t^2)} + \dfrac{R_b^2}{S(b_t^2)} + \cdots \cdots}$$

617. Thus for any piece of work we can now estimate $\dfrac{q_t}{\sqrt{S(q_t^2)}}$. Thus we cannot obtain the ideal mark exactly. But we can obtain each examiner's notion of what the ideal mark should be. Examiner A's ideal mark is $r_a\, q_t$; this is $\dfrac{R_a\, q_t}{\sqrt{S(q_t^2)}}$. Similarly, B's ideal mark is $\dfrac{R_b\, q_t}{\sqrt{S(q_t^2)}}$.

618. We can now write down for each examiner and for each piece of work the ideal mark. We can therefore now obtain by subtraction from the original marks, the random variations. From these random variations we obtain the sums of their squares, which will be better approximations than those calculated. We can also get from each examiner's ideal marks better approximations to R_a, R_b, etc. Thus we can obtain a still better approximation to the ideal marks by substituting these new values in the formula :—

$$\frac{q_t}{\sqrt{S(q_t^2)}} = \frac{x_t \dfrac{R_a}{S(a_t^2)} + y_t \dfrac{R_b}{S(b_t^2)} + \cdots}{\dfrac{R_a^2}{S(a_t^2)} + \dfrac{R_b^2}{S(b_t^2)} + \cdots}$$

[1] In the footnote to para. 611 we saw that $r_a \sqrt{\dfrac{S(q_t^2)}{S(x_t^2)}} = r_{qx}$, so we may write $R_a = r_{qx} \sqrt{S(x_t^2)}$.

In para. 615 we had $S(a_t^2) = S(x_t^2) - R_a^2$, so we may write $S(a_t^2) = S(x_t^2)\,(1 - r^2_{qx})$. Similarly with R_b, R_c ... , $S(b_t^2)$, ...

This formula may therefore be written, taking s_x, s_y, ... , s_q as standard deviations,

$$\frac{q_t}{s_q} = \frac{\dfrac{x_t}{s_x}\dfrac{r_{qx}}{1 - r^2_{qx}} + \dfrac{y_t}{s_y}\dfrac{r_{qy}}{1 - r^2_{qy}} + \cdots .}{\dfrac{r^2_{qx}}{1 - r^2_{qx}} + \dfrac{r^2_{qy}}{1 - r^2_{qy}} + \cdots .}$$

This is in the same form as that used by Prof. Burt in his Memorandum (see Memorandum I, para. 589, p. 299 above), but the formula is not identical with Prof. Burt's. Different arguments have been used in order to arrive at the determination of the ideal mark from the examiners' marks.

Actually it will be found that this process does not need to be carried very far.

619. The new method was applied to the data obtained in the School Certificate Latin investigation from the examiners of Group I. The equations from which approximate values of R_a, R_b, ... are to be obtained are :—

$$R_aR_b = 177{\cdot}4,\ R_aR_e = 173{\cdot}0,\ R_bR_d = 254{\cdot}3,\ R_cR_d = 147{\cdot}0,\ R_dR_e = 238{\cdot}7.$$

$$R_aR_c = 152{\cdot}8,\ R_aR_f = 183{\cdot}8,\ R_bR_e = 196{\cdot}3,\ R_cR_e = 166{\cdot}0,\ R_dR_f = 253{\cdot}3.$$

$$R_aR_d = 253{\cdot}0,\ R_bR_c = 153{\cdot}4,\ R_bR_f = 188{\cdot}5,\ R_cR_f = 156{\cdot}4,\ R_eR_f = 193{\cdot}7.$$

The approximations obtained are as follows :—

$$R_a = 13{\cdot}4,\ R_b = 13{\cdot}9,\ R_c = 10{\cdot}7,\ R_d = 17{\cdot}1,\ R_e = 14{\cdot}0,\ R_f = 14{\cdot}1.$$

We have from the original data :—

$$S(x_t^2) = 209{\cdot}6,\ S(y_t^2) = 232.9,\ S(z_t^2) = 224{\cdot}4,\ S(u_t^2) = 421{\cdot}3,$$
$$S(v_t^2) = 259{\cdot}3,\ S(w_t^2) = 217{\cdot}7.$$

Subtracting from these, R_a^2, R_b^2, ..., respectively, we get approximations to $S(a_t^2)$, $S(b_t^2)$, ..., (see para. 615), as follows :—

$$S(a_t^2) = 29{\cdot}9,\ S(b_t^2) = 38{\cdot}2,\ S(c_t^2) = 109{\cdot}3,\ S(d_t^2) = 130{\cdot}3,$$
$$S(e_t^2) = 62{\cdot}9,\ S(f_t^2) = 19{\cdot}3.$$

620. These approximations to R_a, R_b, ... and $S(a_t^2)$, $S(b_t^2)$, ... enable us to get the ideal marks of each examiner approximately, and from them to get the random variations (see para. 618). From the ideal marks we get the new values of R_a^2, R_b^2, ... as follows :—

$$R_a^2 = 185{\cdot}0,\ R_b^2 = 210{\cdot}3,\ R_c^2 = 118{\cdot}0,\ R_d^2 = 298{\cdot}3,\ R_e^2 = 211{\cdot}1,$$
$$R_f^2 = 214{\cdot}4.$$

We also get the sums of the squares of the random variations :—

$$S(a_t^2) = 28{\cdot}3,\ S(b_t^2) = 37{\cdot}1,\ S(c_t^2) = 98{\cdot}3,\ S(d_t^2) = 88{\cdot}6,$$
$$S(e_t^2) = 62{\cdot}5,\ S(f_t^2) = 14{\cdot}2.$$

These values should be compared with the corresponding figures obtained by calculation (see para. 619 above). Using these new values of $S(a_t^2)$, $S(b_t^2)$, ... together with the appropriate new values of R_a, R_b, ... which are :—

$$R_a = 13{\cdot}6,\ R_b = 14{\cdot}5,\ R_c = 10{\cdot}9,\ R_d = 17{\cdot}3,\ R_e = 14{\cdot}5,\ R_f = 14{\cdot}6,$$

we can apply again the formula for finding the ideal marks (see para. 618).

621. Using these new ideal marks, we get random variations again and obtain these results :—

$$S(a_t^2) = 29{\cdot}3,\ S(b_t^2) = 39{\cdot}9,\ S(c_t^2) = 101{\cdot}0,\ S(d_t^2) = 87{\cdot}3,$$
$$S(e_t^2) = 64{\cdot}3,\ S(f_t^2) = 12{\cdot}6.$$

$$R_a^2 = 177{\cdot}0,\ R_b^2 = 201{\cdot}3,\ R_c^2 = 113{\cdot}3,\ R_d^2 = 284{\cdot}8,\ R_e^2 = 201{\cdot}3,$$
$$R_f^2 = 202{\cdot}9.$$

$$\text{Hence } R_a = 13{\cdot}3,\ R_b = 14{\cdot}2,\ R_c = 10{\cdot}6,\ R_d = 16{\cdot}9,\ R_e = 14{\cdot}2,$$
$$R_f = 14{\cdot}2.$$

These figures should be compared with the corresponding figures in para. 620. It will be seen that there is close agreement.

622. These figures R_a, R_b, ... indicate to what extent there is divergence amongst the examiners in their estimates of the spreading of the ideal marks. We note that *A*, *B*, *E*, and *F* are practically in agreement in the distribution of the ideal marks, but that *C* does not spread the marks to the same extent as these four examiners, and *D* appears to show greater powers of discrimination, judging from the evidence provided by our experimental data. It will be remembered that in the original assumption made in Part II, it was presumed that these values were all equal.

623. The figures R_a, R_b, ... referred to in the previous paragraph may be shown as standard deviations of ideal marks. In the case of examiner *A*, for instance, this standard deviation is $r_a \frac{\sqrt{S(q_t^2)}}{\sqrt{n}} = \frac{R_a}{\sqrt{n}}$, where n is the number of pieces of work. These standard deviations are therefore

(*A*) 3·44, (*B*) 3·66, (*C*) 2·75, (*D*) 4·35, (*E*) 3·66, (*F*) 3·68.

In the original work the standard deviation of the ideal marks was given as 3·76 (see para. 493).

624. The standard deviations of the random variations which emerge as a result of the new analysis are given below, together with those obtained by using the old method. They are

Examiner	A	B	C	D	E	F
New	1·40	1·63	2·59	2·41	2·07	0·92
Old	1·45	1·69	2·66	2·72	2·09	0·88

The differences between the results obtained by the use of the two methods are not material. At any rate the old method, regarded as a first approximation, may be considered to have given results which substantially indicate the relative precision of the various examiners' marking. Naturally the closeness

of these two sets of results is due to the fact that there was not actually a great deal of difference between the examiners' spreading of the ideal marks, due no doubt to the fact that the group of examiners consisted of persons accustomed to the kind of work involved, and well aware of the standards used in this kind of examination.

625. It should perhaps be explained that the results of this analysis which have been just quoted are only intended to refer to the marks obtained in our investigation. We do not pretend to indicate that if these examiners had marked fifteen other scripts they would have marked with exactly the same degree of precision. Our main concern was to find the approximate extent of random variations in such marking, whether it involved 1 or 2 marks, or 5 or 6 marks, or 10 or 11 marks, in our investigation.

626. The new method was also used on the material obtained in the investigation concerned with the differences in marking Essays by Impression and by Details (see paras. 532 *et seq.*). The values of R_a, R_b, etc., proportional to the multipliers r_a, r_b, ... , are as follows :—

Examiner	A	B	C	E	G	K	L	M	N	P	Average	Mean Deviation
Impression	16·2	16·4	9·0	13·3	18·8	13·6	18·6	14·6	15·7	15·1	15·1	2·0
Details	13·3	17·4	11·9	13·3	13·9	14·7	17·1	17·8	15·2	15·4	15·0	1·6

627. We observe that in the case of the Impression marking these multipliers range from 9·0 (*C*) to 18·8 (*G*), that is Examiner *G* spreads the ideal marks twice as much as Examiner *C* spreads them. With the Detailed marking these multipliers range from 11·9 (*C*) to 17·8 (*M*). Thus there is not so much divergence on this account between the examiners when marking by Details as when marking by Impression. This fact is indicated by the mean deviations of the two sets of multipliers. The multipliers in the Detailed marking are on the whole closer to one another than in the marking by Impression.

628. These multipliers, when shown as standard deviations of the ideal marks, are as follows :—

Examiner	A	B	C	E	G	K	L	M	N	P
Impression	15·0	15·1	8·3	12·3	17·4	12·6	17·2	13·5	14·5	13·9
Details	12·2	16·1	11·0	12·2	12·8	13·6	15·8	16·4	14·0	14·2

The standard deviations of the ideal marks obtained by the old method were: Impression 14·4, Details 13·9 (see para. 533).

629. The size of the random variations is indicated below by quoting their standard deviations. They are :—

Examiner		A	B	C	E	G	K	L	M	N	P	Average
Impression	(New)	10·2	9·0	7·1	9·7	11·1	6·7	5·3	7·2	7·8	7·0	8·1
,,	(Old)	10·0	9·0	9·0	9·8	11·5	6·6	6·3	7·3	7·7	7·0	8·4
Details	(New)	8·0	10·7	7·5	9·8	6·4	7·7	7·0	6·3	7·8	6·2	7·7
,,	(Old)	7·7	11·0	7·9	10·0	6·0	8·2	7·2	6·6	7·9	6·3	7·9

(See para. 533.)

The greatest difference between the results obtained by the old and new methods is in the case of Examiner *C* (Impression) where the old figure was 9·0 and the new figure is 7·1. But the general conclusion reached previously, that on the whole no greater precision appears to be achieved when marking by Details than when marking by Impression still stands. Some examiners appear to mark with greater precision by Details than by Impression, e.g., *A*, *G*, *M*, *P* and for some (*B*, *C*, *K*, *L*) the reverse is the case. With *E* and *N* there appears to be no difference. As we stated previously, para. 535, *G* is the only examiner to show a substantial change between marking by Impression and by Details.

630. Thus in this case, where the multipliers introduced by the new method are in a ratio 2 : 1 in the case of a pair of examiners, the results achieved by the original and cruder method described in Part II are substantially accurate. There is only one addition to be made to what was previously said. When marking by Details the examiners appear to keep closer to the Ideal than when marking by Impression (para. 627 above). But it must be remembered that the *order* of the candidates is the same whatever the Ideal marks. On the other hand, the disturbances introduced into the order of merit due to the introduction of the random element in the marking are the same in marking by Details as in marking by Impression.

631. Finally, as a further illustration of the new method, the analysis was applied to the results of the History Honours (Paper II) Investigation (paras. 511 *et seq*). In this case the multipliers introduced by the new method are :—

Examiner	A	B	C	F	H	J	K	L	N	R
	1·2	1·9	1·3	3·2	1·6	1·9	2·1	2·2	4·6	1·1

Here it will be seen that Examiner *N* has a multiplier 4 times that of Examiner *R*. It will be remembered that in the original work these are supposed equal.

632. The ratios of the random variations to the ideal marks

are given below together with similar ratios obtained originally (para. 514).

Examiner	A	B	C	F	H	J	K	L	N	R
New	2·19	1·64	3·54	1·16	3·64	1·89	1·48	1·35	0·97	2·86
Old	2·15	1·44	1·93	1·17	4·31	1·65	1·79	1·61	0·74	1·88

There is some agreement in the results indicated by these figures. Examiner *H* is still indicated as the worst offender by introducing a large element of randomness in his marking; *N* is still the examiner with the greatest relative precision. The orders of these figures are shown below, the smallest being placed first in each case.

Examiner	A	B	C	F	H	J	K	L	N	R
New	7	5	9	2	10	6	4	3	1	8
Old	9	3	8	2	10	5	7	4	1	6

The results achieved by the old method are thus approximately the same as the better results by the new method.

633. Thus even when the multipliers are in the ratio 4 : 1 in the case of one pair of examiners, the old method with its crude results gave quite a good indication of the relative precision of marking of the examiners, judging from the evidence of our investigation.

MEMORANDUM III

ON CERTAIN POINTS OF DIFFICULTY IN CONNECTION WITH SCHOOL CERTIFICATE EXAMINATIONS

By
P. J. Hartog

634. In the Preface to this book attention has been drawn to the fundamental questions of validity and consistency of examination tests (para. (xi) *et seq.*), and incidentally to the way in which these questions affect School Certificate Examinations. The subject is pursued further in this Memorandum, for which most of the material has been derived from the valuable Report of the Panel of Investigators appointed by the Secondary School Examinations Council to enquire into the eight approved School Certificate Examinations held in the summer of 1931.[1] The Panel was presided over by Dr. Cyril Norwood, Chairman of the Council. The Memorandum is not intended as a summary of the Report as a whole, which deals with many other questions besides those referred to here. In what follows we shall designate the Report of the Investigators as the " Report."

635. The question of validity at once raises the question of purpose. The purpose of a School Certificate Examination is defined in Circular 849 of the Board of Education of July, 1914, and subsequent circulars in a number of different formulæ referred to and summed up by the Investigators (Report, p. 11 ; see also p. 147) as follows :—

> " From these quotations it is clear that the primary purpose of the examination was to provide a suitable test of the ordinary work of a Secondary School at the Fifth Form stage, suitable in the sense that whole Forms, and not only picked pupils could properly be presented for it, with the expectation that a large proportion would pass (what proportion was never stated) and that without special preparation or undue disturbance of the normal work of the Form.

[1] *The School Certificate Examination*, etc., H.M. Stationery Office, 1932, 2s. 6d. net.

The secondary purpose of the examination was that it should serve as a qualifying examination for entrance to the Universities and be accepted in lieu of the special examinations for admission to the Professions."

636. The Investigators clearly wish to

"set the School Certificate Examination free from the conditions attaching to matriculation, so that it may serve its primary purpose as essentially a school examination providing an appropriate test of the Secondary School curriculum in its different varieties at the Fifth Form stage." (Report, p. 53.)

We shall limit ourselves in what follows mainly to this "primary purpose."

637. There is, as the Investigators point out, a vague requirement that a "large," but undefined, proportion of the candidates shall be made to pass. And there is a still vaguer requirement that those who pass or obtain credit should have some qualification for passing or obtaining credit; for one of the matters with which the Secondary Schools Examinations Council has to deal is "the maintenance by each approved Examining Body of an adequate standard both for a Pass in the examinations and for a Pass with Credit" (Report, p. 14). The term "adequate standard" is not further defined at this point.

638. A standard of this kind might be defined more or less strictly by some absolute requirement, without reference to any statistical requirements. Such a standard is clearly envisaged by the Investigators in regard to Latin. They report unfavourably on the present tests, and lay down requirements which approach to the ideal of a "utilisable skill" of a modest kind; they recommend that the "result be governed by the actual performance of the candidates and not by the percentages of previous years" (Report, p. 111).[1]

639. On the other hand the Report reveals in many places a conflict between the two kinds of requirement; yet it would appear that, though the Investigators say (with some moderation) that "a pass should mean something in terms of performance"

[1] A prelude to this recommendation reads as follows:—

"The natural result of unsound foundations and scamped work is only too patent to those who inspect the scripts, but it is frequently concealed from the candidates, their teachers and the outside world in general by marking which is extremely lenient and in at least one case by an all-round mechanical addition of marks which makes the whole examination illusory. It is impossible, therefore, to assume that a credit in this subject in the School Certificate Examination necessarily represents a respectable or in some cases even a tolerable level of attainment."

(Report, p. 32),[1] the statistical requirement generally prevails over the standard of performance; and this has important consequences with reference to the question of consistency which we shall point out later.

640. The Investigators again say with reference to the question of purpose in individual subjects :—

> "It is . . . important that those concerned should be in no doubt as to the real purpose of the examination, that the objective in the setting of papers should be clearly defined and the task of Examiners should be freed from all unnecessary complications. Much of the difficulty in securing suitable papers arises indeed from the attempt to make one and the same paper serve diverse and inconsistent ends." (Report, p. 56.)

641. The Investigators give figures to show how the term "a large proportion" in the instructions of the Board of Education has been interpreted. In the eight examinations taken together in the seven years 1925-1931 the average number of passes has only varied from 65·0 per cent. to 69·2 per cent. (Report, p. 30); though there are much wider fluctuations in the percentages of credit and pass in the individual subjects. Nevertheless in the first portion[2] of the passage quoted below the Investigators give what appears to be a statistical definition of the standard for "credit." They say :—

> "The Examining Bodies have been reasonably successful in maintaining a steady credit standard in each main subject. In any given examination the credit mark may vary from subject to subject and is not necessarily the same for a given subject from year to year. The credit standard in a subject in an examination is in fact best described not in terms of the credit mark but broadly speaking as the standard which a given percentage of the candidates offering the subject in question in that examination will reach. . . . The percentage of candidates gaining credit varies considerably in different subjects and in different examinations. Generally speaking, credit is obtained by about half the candidates, but it would be quite untrue, even as a general statement, to say that the credit standard is mechanically fixed to give this result. [But see para. 647 below.] There is a fairly definite relation between the credit and the pass marks in a subject. . . . Of late the Examining Bodies have been endeavouring to keep the pass standard in a subject reasonably uniform from year to year." (*op. cit.*, pp. 31–32.)

[1] In a later chapter the Investigators are somewhat more explicit :—

> "For each paper the crucial question to be answered was 'What, in terms of performance, does a pass or a credit in this subject mean ?' or, more particularly, 'Did the scripts of the candidates who just passed reveal work which was tolerable or was it in fact of little or no worth ?' 'Could the work of the candidates who just got credit be regarded as reasonably good ?'" (Report, p. 55.)

The terms "tolerable" and "reasonably good" are still vague.

[2] i.e., the portion ending with the words "will reach." The second portion seems to be difficult to reconcile with the first portion.

A little lower down follows the statement already quoted :—

"A pass should mean something in terms of performance."

642. It should be pointed out that in an examination which in many respects is to be regarded as a test of the relative progress of the candidates, rather than as a test of their efficiency, there is nothing inherently unreasonable in applying a statistical standard and in saying that nobody who does not reach (say) the 25th percentile will be "passed," and that nobody below the 48th percentile will be awarded "credit."

643. But this at once converts the examination *de facto* into a competitive examination, in which the struggle for marks is serious, since any want of consistency in marking may have serious consequences for individual candidates.

644. It would not appear from the chapters of the Report on special subjects that deserving candidates are as a rule ploughed in masses ; on the contrary the adjustments in some subjects appear, as we have seen, to admit very weak candidates. The question of credits, on which the value of a certificate for Matriculation has hitherto depended, is a different matter.

645. Some of the defects of the present system appear in the view of the Investigators to be due to the setting of papers that are too hard or "rather too hard." A special chapter entitled "Easy Papers and a High Standard of Marking" is devoted to this point.

646. "If," say the Investigators, the papers are "rather too hard, the marking tends to be erratic and the result is unreliable" (*op. cit.*, p. 55). Later, however, they say, "With some notable exceptions the papers set in these examinations cannot be described as difficult" (*op. cit.*, p. 58), and again, "If in these chapters the note of criticism is by no means absent it should be understood that many of the papers and a vast number of questions have been passed over without comment, because they seemed well designed to serve the purpose intended" (*op. cit.*, p. 59). But this general commendation does not appear to modify the criticisms in regard to particular subjects set forth in the extracts which follow.

647. Consider the case of History—taken by the great majority of the candidates. After discussing (*op. cit.*, p. 80) four features common to most of the papers in this subject, the Investigators say (p. 82) :—

"It might be expected that these three common characteristics [out of the four] . . . would have a depressing effect on the average mark and that considerable scaling-up would be required to bring it about that three-quarters of the candidates obtained a pass and just under half a credit. Some wholesale adjustments, either by percentage or by flat-rate additions, had indeed to be made."

648. We have here a fair indication of the statistical requirement in History. In 1931 the total percentage of passes was 77·0, and the percentage passed by the various authorities varied from 70·8 to 79 (*op. cit.*, p. 33). The percentage of credits in History in 1931 varied for the various authorities from 47·1 to 51·6 (*op. cit.*, p. 156). We can gather an idea of the requirements for a pass in terms of performance from the following passage :—

> " The mass of scripts which have been placed before the Investigators affords convincing evidence that by encouraging the reproduction of lifeless text-book formulæ the existing type of examination deadens the pupil's interest in what should be one of the most stimulating subjects in the school curriculum." (*op. cit.*, p. 84.)

649. The Investigators do not say precisely what is the meaning of a " pass " or " credit " in History at the present moment. But from the whole chapter on this subject it may be not unfairly concluded that the aim of the test is vague and hence that the validity is low. In our investigation on School Certificate History scripts, the consistency of the test was so low that it appeared to be difficult to draw any valid conclusion from the marks.[1]

650. In English, taken by over 90 per cent. of the candidates, it is possible, though not easy, to get a closer knowledge of the facts, especially with regard to English Composition. The subject has been dealt with in some detail in the *Essays on Examinations* published by the Committee,[2] so that we can here be brief. The statistical requirement in English seems to be less severe than in History. In 1931 the total percentage of those who passed was 86·2 and the percentage passed by the various authorities varied from 78·6 to 93·1 (*op. cit.*, p. 33). That the standard in terms of performance, as far as English Composition is concerned, is low may be judged from the fact that the Investigators " found much evidence that at present

[1] Reference should be made to the interesting work of Mr. F. C. Happold, who suggested a new type of question in History papers in the number of *History* for July, 1928. A paper on the subject was issued by the Historical Association in 1930 entitled " The Case for Experiment in the Setting of History Papers in the First Schools Examination." In *History* for January, 1932, Mr. Happold gives an account of an experiment carried out by the Delegates of the Oxford Local Examinations on the lines which he suggested. I understand that the Delegates have continued to use Mr. Happold's method is examining Bishop Wordsworth's School, Salisbury, of which he is the Headmaster, and that they have also made other experiments in new types of papers in History. See also *History* for January, 1933, p. 341, and a suggestive and important article by Mr. Happold on " The New History Examinations," in the *Times Educational Supplement* for May 23, 1936.

[2] " English Composition at the School Certificate Examination ; and the ' Write Anything about Something for Anybody ' Theory," by P. J. Hartog, *Essays on Examinations*, Macmillan & Co., 1936, pp. 131-142.

certificates may be granted to candidates whose English is lamentably weak " ; they even mention the word " illiterate." They ask, " *Should a reasonable command of English be required as a condition of obtaining a certificate ?* " ; and they do not venture to answer plainly in the affirmative, the obvious suggestion being that a requirement of that kind would involve a breach of faith with the official instruction that a " large " proportion of the candidates must be passed.

651. The subject of English is particularly interesting from the purely technical, as well as from the general, point of view. It might have been expected that the English test would have been treated as the test of a " utilisable skill," namely, the possession of " a reasonable command of English." If this is one of the chief aims of the examination, there is reason for regarding its validity under present conditions as low. That its consistency is low may be inferred from the remarkable investigation carried out under the auspices of the Durham Board itself (see pp. 64-67 above).

652. From the report of the Investigators in French it may be fairly concluded that the " capacity to translate reasonably well passages of prose of a straightforward kind " (*op. cit.*, p. 103) is a utilisable skill which might well be tested at this examination, but that owing to the attempt to test other things simultaneously, e.g. French Composition, in which " the level of performance reached by the average candidate . . . was in most of the examinations regrettably low " (*op. cit.*, p. 99), and in which candidates, it seems, may now scrape " a few marks on an almost worthless performance " (*op. cit.*, p. 103), the difficulties of marking were great ; and the Investigators say, " the marking of Free Composition was generally unreliable " (*op. cit.*, p. 102). It is interesting to compare these comments, for which no statistical basis is given, with the results of our own investigation (paras. 58-93 above, and especially paras. 92-93). The Investigators regarded the test in French verse as so difficult that they suggest that this should be omitted (*op. cit.*, p. 98).

653. In a paragraph on the efficacy of methods of marking, dealing especially with adjustments at the final stage due to erratic, lenient or overstrict markers, the Investigators write as follows :—

> " In the examinations in German and Spanish, which are far less carefully carried out than in French, these defects are still more serious : many of the papers were so carelessly marked that the results were entirely illusory, and some of the attempts at adjustment at the awarding stage only succeeded in making matters worse." (*op. cit.*, p. 102.)

Yet from the figures for percentages of pass and credit in German (the figures for Spanish are not given) it would seem that the examiners are, at any rate, not greatly influenced by statistical requirements, since the percentages of those who passed in 1931 varied from 50 to 93·9, the total percentage being 71·0 (*op. cit.*, p. 33); and the credit percentages varied from 36·3 to 90·9 (*op. cit.*, p. 156). It will be seen from the Report (p. 24) that the number of candidates who take German and Spanish with the different authorities is very small compared with those who take French. In the opinion of the Investigators, in certain of the examinations the difficulty of the papers set in German and Spanish is such as to discourage schools from presenting candidates in these languages (*op. cit.*, p. 25).

654. In the chapter on Science, the Investigators adopt a different method of defining "credit" from the statistical one quoted in para. 641 above, and more in consonance with that of the chapter on Classics. The following passage is interesting:

> "*Quality and marking of the examination questions.*—It has been emphasised in Chapter VIII that examination papers should be easy and that they should be strictly marked. It cannot be said of all, or even of the majority, of the Science papers set that these conditions are fulfilled. In one instance the paper has been so hard that a score as low as 23 per cent. of the maximum has secured a pass and 39 per cent. the mark of credit. The fact that such low marks have secured these outward signs of success is sometimes concealed by the device of adding a number (or a percentage) of marks to those scored so that it appears that some higher mark has been obtained by the candidates, but the device does not alter the fact that the candidates have not reached the standards which the Examining Body has fixed as its pass or credit line.[1] In such circumstances there is often much uncertainty as to the justice of placing a candidate, or even whole groups of candidates, on one side or the other of a given line." (*op. cit.*, pp. 124-125.)

There is obviously here a conflict between the statistical standard for a pass and any "reasonable" standard of performance.

655. In the same chapter the Investigators also criticise severely the setting of alternative questions (*op. cit.*, p. 129) as implying the measurement of candidates "by as many different 'yardsticks'." Such different "yardsticks" necessarily imply uncertainty for the candidates. Yet in the chapter on History, for reasons which may be conclusive, they suggest that in one paper schools and teachers should, as at present, "enjoy considerable latitude of choice" (*op. cit.*, p. 85).

[1] Compare this use of the word "standard" with its use in the passage quoted in para. 641 above.

656. It is to be noted that the Panel of Investigators consisted of 22 members, a number sufficient no doubt to report on questions of validity in all the subjects dealt with, but insufficient to carry out any adequate tests of consistency on an adequate scale. One or even two additional opinions on scripts bearing the marks of the original examiner cannot be regarded as entirely independent judgments.

657. The Investigators report on this topic (*op. cit.*, p. 37) that " though it cannot be said that considerable variations in the marking of scripts by the several Examiners employed no longer occur in the examinations, great progress has undoubtedly been made in recent years in the standardisation of marking." They admit that " the English Essay is . . . notoriously difficult to mark, and, in assessing the values of answers to questions of the ' essay ' type in History and Geography, there is room for differences of opinion which can only be resolved by consultation among the Examiners and the adoption of an agreed scheme of marking " ; and they continue, " Persons whose interest in School Certificate Examinations is greater than their knowledge of the way in which these examinations are conducted now and again take occasion to point out that if the same scripts are independently marked by a number of different Examiners there will be wide variations in the marking. Examining Bodies who do not allow their Examiners to mark independently are not likely to be greatly moved by this announcement " (*op. cit.*, p. 37).[1] The Investigators do not deal with the fundamental question of possible differences between different Chief Examiners or between different Boards of Examiners to whom the same scripts are submitted. It seems clear that they had not before them numerical data of the kind furnished in the body of this work.

658. We must now refer to those methods of control called " standardisation," in the exercise of which the Examining Bodies are presumed not to " allow their examiners to mark independently." They are described by the Investigators in a special chapter on " The Machinery of Examinations." Some particulars of the method used by the Delegacy for the Oxford Local Examinations are given in a pamphlet by Mr. W. C. Burnet, Secretary to the Delegates, issued in June, 1927, under the heading

[1] In 1931 the total number of Chief Examiners was 251, and that of the Assistant Examiners 1,354 (*op. cit.*, p. 157). In a particular subject there may be 20, 30 or more Assistant Examiners. Although Chief Examiners have, of course, the right to see any scripts corrected by an Assistant Examiner, the vast majority of the scripts corrected by Assistant Examiners cannot, within the limits of time allowed, be revised by Chief Examiners.

"Machinery"; and he has given a further account of it in the *Journal of Education* for January, April, and May, 1936. The method adopted by the Northern Universities Joint Matriculation Board has been described in an excellent book, *Secondary School Examination Statistics*, 1928, by Dr. J. M. Crofts (Secretary of the Board), and Mr. D. Caradog Jones; and also in a later "Statement" entitled "Standardisation" issued in February, 1936, by the Board. After their description of the method of marking the answers, Messrs. Crofts and Jones write :—

> "Now, as a result of all this we have produced a standard which we may say is fairly uniform throughout the panel; we have, in fact, imposed the standard of the chief examiners on the whole panel, but what guarantee have we that this standard is the correct one? What units have the chief examiners to measure by? Actually none whatever. . . .
>
> Take as many fully experienced and trustworthy examiners of the chief examiner type as are available, and let them value independently the same set of answers in the manner described under the paragraph 'Marking the Answers.' Astonishing results will be obtained[1]: every one will expect to get minor variations, but some of the variations obtained will be serious ones. As such experienced and conscientious examiners vary one from another, they cannot all have got the correct absolute standard—one or more may (by accident) have found it, but which? There is no means of telling." (*op. cit.*, pp. 44-45.)

659. A general remedy for the uncertainties of Chief Examiners is suggested in the following passage from the book :—

> "Where large numbers of candidates are being dealt with, the variation of standard among them in the mass from year to year is small, or, at any rate, small compared with the variations which we know take place in the standards of the examiners. The candidates are not like a fruit crop, which may suffer a blight and produce poor results in any one year; in normal times variations in standard are small, and we should err very little if we kept the percentage of passes in the important subjects fairly constant from year to year." (*op. cit.*, p. 45.)

660. In their "Statement" the Board say that in 1935 in the case of 2,697 candidates out of a total which is not given (it was over 17,000 in 1930) there is an agreement with regard to the question of pass and failure (credit is not mentioned) between the expectation of the schools and the awards of the Board in 92·44 per cent. of the cases. But it must be pointed out that "the use of school estimates, and the application of compensation were responsible for 8·5 per cent. of the certificates awarded to the successful candidates in 1935."

[1] No numerical results are given.

The " rank correlation " of the school orders and the examination order is compared in the case of about 2,000 candidates each year. In 1935 the correlations were 0·67 for History, 0·78 for French, 0·71 for Chemistry, and for the aggregate marks 0·88. No figures are given for English. A rank correlation of 0·67 is consistent with very great discrepancies in individual cases ; and it is in the subject of History that our investigations have shown the results to be perhaps the most erratic. It is difficult to explain, without further information, how these figures can be reconciled with the general results recorded above.

661. One of the most interesting features of the " Statement " is a record of an experiment made with photographed scripts, but with only a very small number—from six to eight—which all the Assistant Examiners were asked to mark independently after having each read and valued twenty-five scripts in accordance with the marking-scheme of the Chief Examiners, and having at a meeting all been coached again by the Chief Examiners in the marking of from two to four photographed scripts selected as " difficult." The " Statement " gives a summary of the awards, Failure, Pass, Credit, and Distinction, allotted by the different Assistant Examiners (but omitting the marks). In French Composition (seven scripts) the twenty-four examiners all agreed ; in French Translation (eight scripts) the twenty-four examiners agreed except with regard to one script ; in Chemistry twenty-one examiners agreed in four cases out of six. In History, the awards of the thirty-three examiners were strikingly discrepant. They were as follows :—

Script A	25 Distinctions, 8 Credits.
B	1 Distinction, 29 Credits, 3 Passes.
C	24 Credits, 8 Passes, 1 Fail.
D	33 Distinctions.
E	4 Distinctions, 28 Credits, 1 Pass.
F	20 Passes, 13 Fails.

These results confirm those recorded in Chapter I above.[1]

The " Statement " is unfortunately silent as to both awards and marks in so important a subject as English, taken by nearly every candidate.

662. The " Statement " explains how the Examiners at the end of the examination revalue twenty-five marked scripts of each Assistant Examiner, and describes the measures taken

[1] The average coefficient of rank correlation of the results of each Assistant Examiner with those of the Chief Examiner was 0·984. But Dr. Rhodes has pointed out that there is nothing surprising in this result with six scripts spread out at intervals from fair (Script F) to very good (Script D), with B and E fairly close.

and adjustments made. But the Board has, apparently, not carried out that simple test of uniformity which could be applied to all examiners towards the end of the examination by sending to each one for valuation photographed copies of the same twenty-five, thirty or fifty scripts to mark independently, including some that he has examined before. A crucial test of this kind would show clearly whether or not the standard of the Chief Examiners had in reality been "imposed" on all the Assistant Examiners alike, as it is supposed to be, and would test the consistency of the marking in the best way possible. No other kind of test could afford the same assurance. We suggest that it should be applied in all School Certificate Examinations. The Investigators point out that the Examining Bodies "possess indeed unrivalled opportunities for studying the technique of examining" (*op. cit.*, p. 20).

663. No account of the School Certificate Examinations would be fair without adding that great trouble is taken by all the Examination Authorities to reconsider border-line cases both with regard to Credit and to Pass. Some of the injustices, both to the candidates and to the public (whose rights to a certificate of efficiency are not always recognised in discussions on this subject), that are due to inconsistency in marking are no doubt removed in that way.

664. But, turning now to the results of the investigations recorded in the present book, it seems probable that there are not a few cases where the differences between examiners may far exceed the limits assigned to "border-line cases." Let us for instance consider the investigation in School Certificate Chemistry conducted with two Boards, and elaborate marking-schemes (pp. 51-63 above). Consider Candidate No. 7, who receives from the members of Board I the marks, 48, 47, 42, 37, 31, and 33; and from the members of Board II, 21, 38, 49, 45, 38 and 40 (see Table 29). Here notable differences of opinion survive in spite of all the precautions taken. The extreme range of marks is 28; and the fate of such a candidate in a real examination would be a matter of chance. The case chosen is the extreme one out of thirty candidates. But there are altogether thirteen candidates out of the thirty for whom the range varied from 19 to 28.

665. It is noteworthy that in French (see Table 21, p. 38, above) the ranges were much less, i.e., the consistency of marking was higher than in Chemistry. For the fifty candidates the highest extreme range is only 20; and there are only four candidates for whom the extreme range is 19 or over. Moreover, if

Examiner A's marks were omitted the ranges would be materially diminished. One of the most remarkable features of the investigation on French is the difference between the average percentage marks given by the members of the two Boards to the answers to the same questions, which in four cases amount to 14 per cent., 18·2 per cent., 21·3 per cent., and 23·9 per cent. of the maximum marks allotted in each case to the question (see Table 27, p. 47, above).

666. The comments on School Certificate Examinations made by the authorities to whom we have referred serve to make more intelligible the inconsistencies of examiners recorded in our investigations.

MEMORANDUM IV

A REPLY TO SOME CRITICISMS OF *AN EXAMINATION OF EXAMINATIONS*

By

P. J. Hartog and E. C. Rhodes

667. It has been our purpose to make the present reply to critics of *An Examination of Examinations* as uncontroversial as possible. A mere reference to the original pamphlet or to this book will in our judgment furnish a sufficient answer to not a few criticisms, and these we have not mentioned; but with others we propose to deal briefly below.

668. One not uncommon criticism is that the differences between the markings of the same scripts by independent examiners were "well known." A reference is made in the Preface to the bibliography of the subject. The truth is that in individual cases they could not fail to be "well known," and investigations on a small scale had been made. But no systematic investigation such as the present one had, so far as we are aware, been previously published.[1]

669. Another criticism has been made, chiefly by persons concerned to defend the present "machinery" of large-scale examinations, by which the "crude" marks of a number of Assistant Examiners are manipulated in various ways both by Chief Examiners and by examination authorities with a view to reducing the irregularities of the crude results, as far as possible.

Our critics think that those investigations in which we did not employ "machinery" are "irrelevant" to the present system. Our reply is twofold.

[1] I may perhaps refer here to an address which I gave at the Royal Society of Arts in 1911, when I was intimately concerned with one of the largest and most complex systems of examinations in the country. In that address, reproduced in my book on *Examinations* (Constable, 1918), pp. 30-31, I urged the desirability of a statistical investigation on the independent marking of the same scripts by different examiners as part of a larger enquiry by a Royal Commission.—P.J.H.

(i) *The Examining Bodies do not at present know how far the methods which they use to reduce irregularities are effective.* They can only guess. They have not tested them precisely, as they might have done, by submitting photographic copies of the same set of, say, 30 or 50 scripts (from which all marks of every kind have been removed), *at the conclusion of the examination,* for independent marking by their own Chief Examiners and Assistant Examiners. Such a test would clearly reveal the success or failure of their machinery. It may be asked why it has not been used. (See para. 662.)

(ii) We think it perfectly justifiable to reduce the irregularities of crude results by statistical treatment[1], but it is desirable that the magnitude and nature of these irregularities should be ascertained for each subject in each examination by tests of the kind that we have used. The irregularities may be so large as entirely to discredit the value of the test. We suggest that this is the case with School Certificate History.[2] This is our reason for regarding our investigation on this subject, not only as not irrelevant, but as strictly *ad rem.*

670. The experiment on School Certificate History (pp. 1–16 above), in some ways the most striking in its results, has, no doubt on that account, provoked the most criticism. The object of that investigation was to ascertain what marks would be allotted to a set of scripts which had all been assigned the same marks by an examination authority if no external standard were "imposed" on the examiners by a Chief Examiner; and to ascertain further how the judgment of the examiners would vary when they marked the same scripts a year or more later.

671. We may summarise the results by saying that on the first occasion, instead of the same "middling mark" of the examining body, the fifteen examiners of the fifteen scripts allotted 42 different marks varying from 21 to 70; and on the second occasion (one examiner being unable to serve again) the fourteen examiners allotted 44 different marks varying from 16 to 71; further, with regard to the 210 verdicts of Failure, Pass and Credit, the fourteen examiners changed their minds in 92 cases. The average range of the unadjusted marks for the various candidates was 25 on the first occasion and 26 on the second.

[1] For an example of the use of such methods see Part II, and Memoranda I and II of this book.

[2] See also with reference to this subject, Memorandum III, paras. 647–649, 661 and 666 above.

672. What are the chief criticisms ? The first is that, as the scripts " were all of equal merit," the examiners were " misled " when we asked them to allot to the scripts awards of Failure, Pass and Credit, as well as numerical marks, and that they naturally assumed that the marks should be distributed over a wide field. We ought, say some critics, to have told the examiners " that the scripts were approximately of equal merit."

673. To this the reply is simple. To have given the examiners any information on this point would have invalidated the investigation. We regard the evidence, including the assignment of marks varying from about 20 to 70 to specimens of these scripts on two occasions, as conclusive against any assumption *a priori* that the papers were actually of equal merit, so that, as the event proved, to have made the assumption at that stage would have been absurd. It was not made by us, and with its disappearance the criticisms on this score vanish also.

674. In our view, the results are so discrepant that it is difficult to draw any firm conclusion from them. Although there is some correlation between the orders of the " ideal " marks of the two investigations (see Table 120, p. 201) it is, nevertheless, just conceivable that the verdict of the original examining authority might be the right one. But apart from such minor adjustments as the investigations on " ideal marks " (paras. 430–433) might suggest, we see no particular reason for regarding the verdict of the original examining authority as either superior or inferior to any of the 29 other sets of verdicts.

675. The group of candidates was probably mediocre in the sense that there were no scripts adjudged by any of the examiners to be of outstanding merit, and none adjudged to be valueless. But the remarkable differences between the individual changes in the verdicts of the individual examiners, passed over in silence by nearly all our critics, have convinced us that the inconsistency of the marking indicates not a failure of any individual examiner or of the original examining authority, but a failure in method.

676. The scathing condemnation of the Investigators on School Certificate History scripts in general, if justified (see Memorandum III, para. 648), indicates the want of validity of a test of this kind. Our own investigation indicates its want of consistency, confirmed by the breakdown of the similar and carefully prepared test in consistency on a small scale carried out by the Northern Universities Joint Matriculation Board (Memorandum III, para. 661).

677. A further and interesting criticism is that an examiner

in History cannot "get a standard" for an award without correcting, say, fifty scripts. Such a statement shows how far removed the test must be from a test of a utilisable skill (see Preface, paras. xviii–xxii, and Memorandum III). The order of merit in which the scripts were placed by the different examiners at the first investigation shows extraordinary differences of opinion. These differences were not less on the second occasion, though the individual examiners varied their own orders considerably. The investigation of the orders of merit negatives the suggestion that had the number of papers been multiplied the results would have been more regular.

678. Our conclusion as to the facts is unaltered. It is urgent to devise a system of tests in History that shall be both more valid and more consistent. It is for historians to try out such methods (see para. 649, footnote 1). The criticisms have in no wise weakened our conviction that our Committee were right in saying that the element of chance in the School Certificate History Examination, under the present system, is gravely disturbing.

679. *School Certificate French scripts* (pp. 34–50 and 204–210). One critic has suggested that the French examiners (see pp. 34–50) were unconscientious in carrying out their directions. The fact that the Chief Examiners, after examining eight or nine of the fifty corrected scripts in each case, found no such fault may be regarded as conclusive evidence to the contrary; and the slight adjustments they made in the final marks (see para. 67) show that there was nothing abnormal in the method of marking.

680. *History Honours scripts* (pp. 152–167 and 231–234). Two criticisms have been made of this investigation. The first is that the examiners could not "get a standard" with so few scripts as 16 or 18. If this be so there must be a very large number of University examiners in many Honours subjects who boldly arrange candidates in order of merit and in classes without ever "getting a standard."[1]

Other reasons for regarding the objection as invalid have been given in para. 677 above. It is, we think, fair to assume that even without being able to "get a standard" for classes the

[1] The number of successful candidates for seven of the latest Honours degree examinations at Oxford taken from the Calendar for 1936 were as follows: 19, 14, 1, 13, 10, 1, 16. The corresponding figures for nine of the latest Tripos examinations at Cambridge were 18, 2, 1, 6, 3, 5, 9, 6, 11. It is of course true that some candidates may have failed or got a "pass"; but with a Fourth Class at Oxford and a Third at Cambridge, we know that the numbers are small. Unfortunately the Calendars of the newer universities which we have consulted do not give the corresponding figures.

individual examiners felt no particular difficulty in arranging the candidates in order of merit to their own satisfaction.

On p. 232 we give the correlation coefficients between the orders of the different examiners for each of the Papers, which are summarised below :—

	Minimum	Maximum	Average
Paper I ..	·31	·64	·46
Paper II ..	—·10	·59	·26
Paper II ..	—·04	·85	·45
Paper IV ..	—·05	·73	·45

We may justifiably say that this is a collection of low correlation coefficients.

681. The second objection is twofold : (*a*) that we did not get the examiners to agree at the outset and at a formal meeting to a common interpretation of the symbols $\alpha +$ to δ by means of verbal formulæ showing what qualities (and defects, if any) each symbol should connote ; and (*b*) that we did not ask examiners to agree at the outset where the border-lines between the different classes should be placed in terms of the symbols. The assumption made by our critics is that if examiners had agreed in this way the symbols assigned would have " meant the same thing."

682. The precautions that we took are described below.

The following query was sent out with our marking-scheme ($\alpha +$ to δ) to all the examiners except two who consented to act at too late a stage for an alteration to be made : " Will you kindly let me know if the scheme submitted to you appears satisfactory or, if you regard it as defective in any respect, will you kindly suggest such amendment as seems to you desirable ? " The examiners all assented to the scheme, thus confirming the view of the eminent authority by whom we were guided throughout in this matter that it would be understood by all our History Honours examiners, and that presumably they would " mean the same things " when they used the same symbols. As is shown, there were gaps in our list of symbols between $\beta\alpha$ and $\beta++$, and between $\beta=$ and $\beta\gamma$ which we interpreted as corresponding to class divisions. Only two examiners made comments on these gaps in the first instance. Not a single examiner suggested the desirability of a formal meeting to discuss the " meaning " of the symbols before the marking was begun ; and our adviser has authorised us to say that, though he could remember occasions on which the examiners had decided at what points the division into classes should be made, he could assure us that, so far as his extensive experience

went, the existence of a common understanding on the value to be attached to the symbols was taken for granted. It was past experience of a common tradition that mattered, not minute (and in his mind impossible) agreement on the exact meaning to be attached to symbols.[1] After the completion of the marking, in order to make assurance doubly sure as to this interpretation of class limits, we wrote to all the examiners to ascertain their individual interpretations of the symbols in terms of classes and received the varying replies summarised in *An Examination of Examinations*, pp. 71 and 72, and in this book on p. 163.

683. The following examples will show that even when two examiners are in close agreement with regard to the symbols which should determine the limits of different classes they may agree exactly as to the mark to be assigned to some scripts and differ fundamentally as to the mark to be assigned to others.

Examiners H and N agree very closely in their class limits ;[2] yet for Paper II they give the following awards to three different candidates (see pp. 157 and 158) :—

		Examiner H			Examiner N		
Candidate No.	1	$\beta\alpha$	(16)	2nd Class	$\gamma\beta$	(5)	3rd Class
,,	13	$\beta?+$	(12)	2nd Class	$\beta?+$	(12)	2nd Class
,,	14	$\gamma\beta$	(5)	3rd Class	$\beta\alpha$	(16)	2nd Class

If H and N " meant the same thing " when they assigned the same mark and class to Candidate No. 13, what did they mean when they assigned different marks and classes to Candidates Nos. 1 and 14 ?

We give another instance relating to Paper II. A and C agree very nearly, though not quite so closely as H and N, as to their class limits. We quote in each case, as before, the awards in literal and numerical marks and in classes assigned by the examiners in question for four candidates :—

		Examiner A			Examiner C		
Candidate No.	3	β	(11)	2nd Class	β	(11)	2nd Class
,,	8	β	(11)	2nd Class	β	(11)	2nd Class
,,	4	$\beta-$	(9)	2nd Class	$\alpha\beta$	(17)	1st Class
,,	9	$\alpha=$	(18)	1st Class	$\beta-$	(9)	2nd Class

If A and C " meant the same thing " when by their symbols they assigned the same mark and class to Candidates Nos. 3

[1] After the publication of our pamphlet one examiner made an anonymous statement that in his view there should have been a meeting of the examiners to discuss the meaning of the symbols.

[2] This is more manifest in the detailed original documents than in the summary on p. 163.

and 8, what did they "mean" when they assigned such different marks to Candidates Nos. 4 and 9 ?[1]

Again, for Paper II the following are the marks of Candidates Nos. 7 and 8 :—

Candidate	Examiner A	B	C	F	H
No. 7 ..	βα (16)	αβ (17)	β+ (13)	β? − (10)	βγ (6)
No. 8 ..	β (11)	α − (20)	β (11)	β? − (10)	β − (9)

Candidate	Examiner J	K	L	N	R
No. 7 ..	β − (9)	γ (3)	β?+(12)	β (11)	β?+ (12)
No. 8 ..	γ+ (4)	β+ (13)	β − (9)	β+?+ (14)	β++(15)

Four examiners, A, C, J, and L, regard No. 7 as superior to No. 8. Five examiners, B, H, K, N, and R, regard him as inferior, and one of these, K, as inferior by ten grades. One examiner, F, regards the two as equal.

No juggling with verbal formulæ can conceal the differences of opinion of the examiners in such cases. Our critics do not appear to have thrown any new light on the subject.[2]

As indicated previously, during the whole procedure from beginning to end, we acted in accordance with the advice of an historian of the most unquestioned experience and rank.

684. We have to add that our purpose was not to reproduce the exact conditions of an examination in History Honours, but to ascertain the degree of variation between experienced examiners in History judging the same scripts. In many examinations in History the awards are greatly affected by the previous work of the students and the judgments of their teachers; the examination may be a subordinate factor in an award. Again, at examiners' meetings there may be all

[1] A similar instance of the same kind of differences may be quoted with regard to Paper IV (see pp. 161 and 162) on which a third pair of examiners, whose class limits are almost the same, gave the following awards :—

	Examiner A	Examiner L
Candidate No. 3	β+ (13) 2nd Class	β/+ (12) 2nd Class
,, 8	βγ (6) 3rd Class	β/+ (12) 2nd Class
,, 9	α (22) 1st Class	β+ (13) 2nd Class

Thus, while Examiner L regards the three candidates as almost equal, A agrees with him about Candidate No. 3, but regards Candidate No. 8 as sixteen grades below Candidate No. 9. Other similar cases, but with less well-marked differences, can easily be gathered from the relevant Tables.

[2] One critic has suggested that at this point further investigation of the reasons both for agreements and differences of examiners with regard to sets of scripts such as those referred to would be of value. We concur. But the investigation would necessarily involve much time and considerable expenditure on publication. It should be carried out by specialists in each subject and in experimental psychology. The present results indicate a number of fields in which such investigation is desirable.

kinds of compromises and the judgment of certain scripts may be modified by a *viva voce*. On the other hand, there are important examinations in History in which the fate of candidates is determined on written work by two examiners or only by a single examiner. How large the element of chance must be in such examinations is shown above.

685. *Viva Voce Investigation* (see pp. 168–176). In dealing with our investigation on a Viva Voce Examination, one critic suggests that we have neglected the variability of the candidates.[1] It is difficult to see how the variability both of candidates and of examiners could have been further reduced. The candidates were examined for two periods of only from a quarter of an hour to half an hour, separated in most cases by a luncheon interval. They were provided with a comfortable common-room, newspapers, etc. It is difficult to see what further precaution could be taken. The suggestion that the candidates were seriously affected by their first examination, either by being unduly exhilarated or unduly depressed, seems to one of us who was present during the whole time at one or other of the examinations to be gratuitous (see p. 176 above). It is conceivable that the variability of candidates (and of examiners) may be so great as entirely to invalidate any *viva voce* test of fitness for a job, however well conducted, though this is at present contrary to general belief. We made no assumption on this point, positive or negative.

686. Finally, we have been reproached for directing the attention of our readers in some cases rather to the differences between the marks of different examiners than to the agreements. That is true: it is the differences between the marks that indicate the weak points in the system which need to be strengthened.

[1] See on this point Hartog on *Examinations*, p. 22, and the article on Examinations by Hartog and Watson in the Encyclopædia Britannica, 11th Edn., p. 48.

Printed at The Chapel River Press, Andover, Hants