INTERNATIONAL INSTITUTE EXAMINATIONS ENQUIRY

AN EXAMINATION OF EXAMINATIONS

Being a Summary of Investigations on the 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

SECOND EDITION

Fourth Impression

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

Price One Shilling (By Post 1s. 2d.)

AN EXAMINATION OF EXAMINATIONS

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.

INTERNATIONAL INSTITUTE EXAMINATIONS ENQUIRY

AN EXAMINATION OF EXAMINATIONS

Being a Summary of Investigations on the 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

SECOND EDITION

Fourth Impression

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

WATERLOW and SONS LIMITED,
LONDON and DUNSTABLE.

TABLE OF CONTENTS

PREFACE TO THE FIRST EDITION

1. 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.

2. 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.

3. The investigations on examinations of which this pamphlet is a summary 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

[1] In a Conspectus in preparation by the Committee there will appear between 150 and 200 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.

Conference committees were set up in all the European countries above-named. Each of these committees received a grant for 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. This pamphlet is substantially identical with the report presented by the English Committee to the Folkestone Conference, and it is published in its present form in accordance with a wish expressed at that Conference.

4. 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 C. W. Valentine joined the English Committee in the present year[3]. The address of the English Committee is 1, Plowden Buildings, Temple, London, E.C.4.

Dr. Henry Suzzallo, President of the Carnegie Foundation, New York City.

Dr. Edward L. Thorndike, Professor of Education, Teachers College, Columbia University.

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

FRANCE—

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

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

M. Bouglé, Directeur-adjoint de l'Ecole 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,

5. The Committee engaged Dr. E. C. Rhodes, Reader in Statistics in the University of London, to act as their statistician.

6. 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 Examina-*

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.

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'Education, 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'Ecoles, Troinex, Genève (Dr. Soc.).

Dr. Charles Junod.

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

M. Jean Piaget, Directeur du Bureau International d'Education.

tions (1900-32)[4], which shows how much has been written on the subject in this country during the first third of the century. The Committee are also publishing a volume of *Essays on Examinations*, dealing with a number of aspects of the subject, which will appear soon after this pamphlet, and a *Conspectus of Examinations in Great Britain and Northern Ireland*, which will appear later. But the main work carried out for the Committee will be recorded in a volume entitled *The Marks of Examiners*, now in course of printing, of which the present pamphlet is a summary.

7. The object of the investigations to be described 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 piece of Latin prose varied from 45 to 100 per cent. In the United States, Messrs. Starch and Elliott, and, in France, M. Laugier and Mlle. Weinberg have found similar results, but no systematic comparison has hitherto been published of the marks allotted by a number of different examiners, all experienced and qualified for their task, to sets of scripts (answer-books) actually written at public examinations. Both the English and the French Committees have attacked this subject, and the present pamphlet gives a fairly extended summary of the English results and a brief one of the French. These results are similar in the two countries, and equally disquieting. It is clear that the part played by chance in the verdicts given at different examinations on which careers depend must often at the present moment be a great one. The Committee are well aware that the consideration of borderline cases by examination authorities does materially diminish the chances of a candidate being wrongly rejected; but it must be pointed out that candidates may be placed in error below the

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.

[4] *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.

"borderline." Again, it must be remembered in the interest of the public, to whom an examination certificate means a certificate of efficiency, that candidates may now by chance obtain such certificates when they should by rights be rejected.

8. Of all the results recorded by the English Committee perhaps the most disturbing are those recorded in the investigation on the marking of School Certificate History scripts. It was found that when fourteen experienced examiners re-marked independently fifteen scripts which had all received the same moderate mark from the examining authority by which they were furnished, these examiners, between them, allotted over forty different marks to the several scripts. It was found, further, that when these examiners re-marked once more the same scripts after intervals of from twelve to nineteen months, they changed their minds as to the verdict of Pass, Fail, and Credit in 92 cases out of the total of 210. Clearly a test of this kind cannot inspire confidence.

9. Our investigations show that the employment of boards of examiners instead of individual examiners, though it diminishes, does not remove the element of chance in examinations, and that boards, as well as individuals, may disagree in their verdicts. The element of chance in examinations still subsists to a dangerous degree in the subjects which have been investigated by the Committee.

10. The question may at once be asked: Should examinations be abolished? If not, what remedies can be suggested?

The Committee are clearly opposed to the root and branch policy. They are of opinion that examinations as a test of efficiency are necessary. They are further of opinion that, in addition to those examinations which yield identical results when applied by different examiners (*e.g.* "New Type" or "Objective" examinations), the traditional "essay" examination should be preserved. 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. No doubt investigations like those recorded by our 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.

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.

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.

PREFACE TO THE SECOND EDITION

The first edition of this pamphlet appeared in December, 1935. A few slips were corrected in the second impression, which was issued shortly afterwards; and some further corrections of detail have been made in the present edition. These corrections do not in any way affect the conclusions of the Committee. While the pamphlet has been received with warm approval by the general public, it has evoked certain criticisms, with some of which it is proposed to deal in *The Marks of Examiners*, now nearly ready for publication. To have dealt with the criticisms in this pamphlet would have involved an increase in both its size and price which was thought undesirable.

It should be added that Professor Valentine, who was elected a member of the Committee in July, 1935, resigned at the end of December in the same year, and that Professor F. Clarke, M.A., Advisor to the Overseas Students in the Institute of Education of the University of London and Director-elect of the Institute, who has been in close touch with the Committee for some time, became a member early in 1936.

April, 1936.

Part I—General

Introduction

1. The main object of the investigations was to test the concurrence of the marking of a number of examination scripts by a number of independent examiners, or, in certain cases, by two independent boards of examiners.

2. In carrying out the investigations, the following general principles were observed :—

(i) 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.

(ii) The following examinations were selected by the Committee for the purpose of the investigations, 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 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 at the age of between 10 and 12, on the results of which children in elementary schools 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.*

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

(iv) 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 the panel of a single examining body (other than the body which had supplied the scripts).[1] The examiners for the College Entrance Scholarship Essay scripts and for the University Mathematical Honours scripts were in either case 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 17 examiners concerned were chosen from nine different universities and included nine university professors.

(v) 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.

(vi) 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.

(vii) 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.

(viii) The marks were all analysed by Dr. E. C. Rhodes, Reader in Statistics in the University of London, and the results have been prepared for publication by the Director and Dr. Rhodes

[1] In the investigation on School Certificate English conducted under the auspices of the Durham University School Examinations Board, of which we are printing and extending the results, the examiners were not all chosen from the panels of the same examining body.

and submitted to the Committee. The volume containing the details of the investigations will extend to about 250 pages, and will comprise two sections: Section I, containing the important details and figures for each investigation, and Section II, containing a more elaborate statistical analysis by Dr. Rhodes, in which it is attempted to separate the differences of marking due to difference of the standards adopted by the individual examiners from the random deviations of each examiner from his own standard. It will include additional memoranda by Professor Cyril Burt and Dr. Rhodes on the most suitable methods of analysis for data of this kind.

(ix) 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.

3. 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 all well qualified, would tend to affect the fate of individual candidates.

4. Besides the investigations into written examinations, the Committee carried out one investigation of a particularly interesting kind into the concurrence 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 defined in para. 81(*d*) below.

The results of the different investigations are briefly summarised in the following sections.

School Certificate History

5. Fifteen scripts were selected which had been awarded exactly the same "middling" mark by the School Certificate authority concerned, and these scripts were marked in turn and independently by 15 examiners, who were asked to assign to them both marks and awards of Failure, Pass and Credit. After an interval which varied with the different examiners, but was not less than 12 nor more than 19 months in any instance, the same scripts, after being renumbered, were marked again by 14 out of the 15 original examiners (one examiner being unable to serve again).

The 14 examiners assured us that they had kept no record of their previous work and this was indeed obvious from the results.

6. Whereas the scripts had been all allotted the same moderate mark by the original examining body, they were allotted by the 15 examiners on the first occasion 42 different marks out of a maximum of 96, varying from 21 to 70. On the second occasion the total number of the different marks was 44, and the marks varied from 16 to 71. There is no space here to analyse the differences of the marks allotted by the various examiners to the same candidates. In one case the difference was 30 marks out of the maximum of 96.

7. Perhaps the most striking feature in the investigation is this: On each occasion the examiners awarded not only numerical marks, but the verdict of Failure, Pass or Credit. In comparing the two sets of awards we can only take into account the 14 examiners who acted on both occasions. On each occasion the 14 examiners awarded a total of 210 verdicts to the 15 candidates. It was found that in 92 cases out of the 210 the individual examiners gave a different verdict on the second occasion from the verdict awarded on the first.

8. In nine cases candidates were moved two classes up or down. One examiner changed his verdict in regard to eight candidates out of the fifteen. Yet he only varied his average by a unit, and he awarded the same number of Failure marks, one less Pass, and one more Credit. Such irregularity of judgment is not only formidable, but it is one which would not be detected by any ordinary analysis. Statistically his results on the two occasions were almost the same, but the fate he allotted to half the candidates was different.

In some cases the examiners altered their general standard on the second occasion. One examiner moved 8 candidates down a class, and one down two classes. Another examiner moved 7 candidates down a class. Of the 14 examiners there is only one who was exceptionally steady and whose numerical mark never varied by more than 7 out of 100.

9. It may well be asked, in view of the extreme differences of these results, what validity can be attached to the marking of School Certificate History papers. It is perfectly true that, as Professor Spearman has pointed out, validity and "reliability" or concurrence of marking are by no means equivalent terms.

but no process of measurement can be valid when it yields such discrepant results in the hands of the same examiners on two different occasions.

School Certificate Latin

10. This investigation dealt with two 2-hour papers, of which the marks were added together. The scripts of 15 candidates were so selected that the candidates had obtained at the original examination exactly the same moderate mark for the two papers combined. 15 examiners were appointed, of whom two were treated as Chief Examiners for the drafting of a marking-scheme. The examiners were furnished with examination papers (though not with "trial-scripts" as in later experiments). The marking scheme was finally settled after correspondence with all the examiners concerned on all points regarded as contentious. The correspondence showed that six of the examiners preferred more detailed instructions in respect of unprepared passages than the other seven, and it was decided to adopt two marking-schemes to meet the wishes of the different examiners concerned. The examiners were therefore divided into two Groups—Group I, consisting of six examiners who used Scheme I, and Group II, consisting of seven examiners who used Scheme II. The two schemes differed only by the addition of 19 more detailed instructions in respect of unprepared passages from and into Latin in one paper than the other. Of these, 10 were allotted to a question which was only selected by a single candidate. The maximum for each question and the total maximum were the same in the two Schemes. It is obvious that the two Groups cannot strictly be regarded as analogous to two independent Boards, who would no doubt have adopted marking-schemes differing far more widely.

11. Whereas the fifteen couples of scripts had originally been assigned the same moderate mark, under Scheme I they received from the 6 examiners concerned 24 different marks ranging from 28 to 55; and under Scheme II they received from the seven examiners concerned 28 different marks ranging from 33 to 61. The total number of different marks allotted under the two schemes was 31 and the total range from 28 to 61. It is quite obvious that in spite of the detailed marking schemes the individual examiners adopted very different standards.

12. A detailed analysis has been made of the marks for the different questions. These questions were originally marked

on a higher scale, which was reduced so as to yield a maximum for the two papers of 100. It is remarkable that the difference between examiners varies very much with the candidate. Thus for one candidate the marks for a question of which the original maximum was 60 (translation from Cæsar), the extreme range of the marks allotted by the 13 examiners is only 9 marks, whereas for another candidate the extreme difference was 28 marks or 47 per cent. of the maximum. In the case of some questions on accidence the difference between the marks is very small.

School Certificate French

13. The scripts investigated were written as answers to two 2-hour papers. Two independent Boards were set up, each consisting of a Chief Examiner and six other examiners. The examining body supplied at our request 150 scripts altogether, chosen so that the marks allotted by the original examiners corresponded to a normal frequency distribution and ranged from the worst to the best. Of these 50 were selected, corresponding to the same normal distribution, for final marking, and were reproduced photographically. The others served as "trial-scripts."

14. Each Chief Examiner drew up his own marking-scheme, discussed it with his Board in the ordinary way and, after settling his scheme, gave each of his Board a number of trial-scripts to mark so as to control the methods of marking of each examiner. As a result of this process the two Boards quite independently adopted complex schemes, which were, however, obviously the result of a common tradition. Board I gave 5 general directions, and 640 detailed directions for Paper I and 290 detailed directions for Paper II, mainly concerning points of English and French in translation. The scheme of Board II included 700 detailed items for Paper I and 300 for Paper II. These detailed directions did not require any appreciable effort of memory on the part of the examiners. Although 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 50 cases were actually conflicting. Each Board settled its own standard for Failure, Pass or Credit. The Chief Examiner, after seeing samples of the trial markings of each examiner, gave instructions for his marks to be raised or lowered in some particular way.

15. The returns of the individual examiners showed that the number of Failures varied from 6 to 15, of Passes from 7 to 16, of Credits from 21 to 30, and of Distinctions from 1 to 9. Agreement was reached between the 6 examiners of Board I on the awards to only 27 candidates out of 50, and agreement was reached between the examiners of Board II in regard to only 30 out of 50. The average range (the difference between the highest and lowest mark allotted by the different examiners to the same script) for Board I was 10.6 marks and for Board II 7.8, out of one hundred. The extreme range was 19 for Board I, and 16 for Board II.

16. One of the interesting features of the marking of the two Boards was that the average mark of Board I for a piece of dictation expressed as a fraction of the maximum was 14 per cent. higher than the corresponding average mark of Board II, and that the average mark of Board I for a question involving translation from English into French, expressed as a fraction of the maximum, was about 24 per cent. lower than the corresponding average of Board II. The maxima were approximately the same for the two Boards.

When we consider the average marks for the two Boards of the scripts treated as a whole, such differences disappear; but the fate of individual candidates depends on these differences which a similarity of general results effectively conceals. A candidate who did poorly in dictation would be more leniently treated by the examiners of Board I. A candidate who did poorly in translation from English into French would be more leniently treated by the examiners of Board II. Moreover, the fate of a candidate might depend on the particular member of the Board to whom his script is assigned for marking.

School Certificate Chemistry

17. The procedure in the case of Chemistry was almost identical with that adopted in the case of French, but the number of final scripts selected for the final marking was only 30 instead of 50, as the average length of the scripts was considerable.

Board I in its marking-scheme gave about 95 detailed directions to the examiners, and Board II about 85. For certain details the two Boards gave the same marks, for others they gave marks appreciably differing. The differences between the Boards would no doubt have been greater but for the fact that the candidates were instructed to select any six questions out of

eight, so that it was necessary to allot identical or almost identical maxima to the different questions.

18. In the returns of the individual examiners of the two Boards, taken together, the number of awards of Failure varied from 5 to 10, of Passes from 2 to 11, of Credit from 9 to 16, and of Distinction from 0 to 8. No mere adjustment of averages would remove such discrepancies between the distributions of awards by individual examiners. The differences between the two Boards in respect of different questions is less than in the case of French, but for one question, dealing with a simple question of chemical theory, the average mark for Board I was 33 per cent. of the maximum, while the corresponding average for Board II was 46. It is only in regard to this point that we get anything comparable to the remarkable differences which were found between the two French Boards (*see* para. 16 above). Nevertheless it is true that, as in French, the fate of a candidate depends very largely on the personnel of the Board, and on the particular examiner to whom his script is assigned. The average range of marks was 10 for Board I, and 10.9 for Board II, out of one hundred. The extreme range was 25 for Board I, and 28 for Board II.

School Certificate English

19. We include in this Report details of an investigation on School Certificate English, carried out just before our own work was begun, on behalf of the Durham University Examinations Board. It was on lines similar to those which we have adopted, and yielded similar results. An analysis of the figures by Mr. C. Roberts and Professor H. V. A. Briscoe was published by permission of the Durham Board (in *The A.M.A.* for Dec. 1931, and Feb. 1932). The detailed mark-sheets were later furnished to us by the Board, and we have made use of these in both parts of this Report. The whole of the English scripts from one school, 48 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.

20. The examiners all accepted the marking-scheme of the Chief Examiner of the Durham Board.

21. There were two papers: Paper I, a 2 hours' paper on Essay and Précis, and Paper II, a 3 hours' paper, mainly on set books in prose and verse. The marks for the two papers were added and then reduced so as to correspond with a maximum of 100.

22. The minimum range, *i.e.*, the extreme difference between the marks allotted to an individual candidate, was 7, the maximum 31, and the average 18·5. But the differences between the examiners was shown most clearly by the differences between the award of Failures, Passes, Credits, and Special Credits of the individual examiners.

The following Table shows the numbers of awards:—

Examiner	*Fail*	*Pass*	*Credit*	*Special Credit*
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

23. An inspection of the figures in greater detail shows that in the case of only *one* candidate out of the 48 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 examiners were in agreement. Examiner G "ploughed" 19 candidates, while no other examiner "ploughed" more than seven, and two "ploughed" none; Examiner B awarded 12 "Special Credits," while the other examiners awarded very few or none.

24. The examination is not a competitive examination, and therefore the order of merit is not of any special importance to the candidates. But the differences of opinion of the different examiners in regard to their relative merits are shown by the following statement:—

The difference between the highest and lowest position assigned to a candidate is—

30 or more in 5 cases
20—29 in 19 cases
10—19 in 18 cases
Under 10 in 6 cases

25. The divergencies of the marks allotted to the two Papers considered separately were greater than those shown when the marks were added together.

26. Mr. Roberts and Professor Briscoe draw attention to certain extreme divergencies. On 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 39 candidates who were awarded a Failure mark by one or more examiners, 25 were awarded a Credit, 8 Special Credit, and 3 Distinction by one or more examiners. Again, 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.

27. 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

In Paper II, again, 36 of the 48 candidates were passed by all seven examiners.[1] Of the remainder 3 received a Failure mark from only one examiner, and 8 by from 2 to 4 examiners; but in all these cases the candidates were awarded from one to three Credits by the 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.

[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 Composition at the School Certificate Examination by Sir Philip Hartog in the "Essays on Examinations" published by the Committee.

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.

28. The following Table shows the numbers of awards of the different examiners on Papers I and II separately :—

	PAPER I					*PAPER II*				
Examiner	*Failure*	*Pass*	*Credit*	*Special Credit*	*Distinction*	*Failure*	*Pass*	*Credit*	*Special Credit*	*Distinction*
A	7	15	15	7	4	2	12	30	4	0
B	7	10	23	5	3	0	2	27	15	4
C	12	29	7	0	0	10	21	17	0	0
D	9	17	20	2	0	1	7	25	15	0
E	8	20	18	2	0	6	17	23	2	0
F	6	21	17	4	0	2	6	36	4	0
G	35	11	2	0	0	9	16	18	4	1

It is to be remembered that Examiners B and F ordinarily examine for one Examining Body, and Examiners C, D and E ordinarily examine for another Examining Body.

29. We believe that the method of selection of examiners for our investigations was such as to enable us to draw general conclusions from our results. The independent investigation carried out by the Durham University Board yields valuable support to our conclusions.

Special Place Examination (I) : *Arithmetic and English*

30. This was the most complex of all the investigations, since it dealt with two subjects. The scripts of 150 candidates in Arithmetic and in English were marked by 10 examiners in each subject. The marking-schemes were settled after correspondence with the examiners, each of whom marked 50 trial-scripts in accordance with a draft marking-scheme before expressing his opinion on the scheme. The marking-schemes were modified in such a way as to deal with all the points raised by the individual examiners, and they were finally settled only after an assurance had been received from each of the examiners in the subject concerned that the schemes contained no ambiguities.

31. The 150 scripts for the final investigation included a large proportion of the very best sent in for the original examination,

as judged by the original examining authority. A very high proportion of these scripts would therefore be scripts of successful candidates and of those who approached success.

32. The results were first analysed in the following way: At the original examination the fate of a candidate would primarily depend on the marking of a couple of examiners, one for English and one for Arithmetic. Of the examiners actually employed, couples were chosen at random and designated A, B, C, D, etc. As an example of the differences of marks of these couples, we may choose Candidate No. 1, who received from the 10 couples of examiners the following marks out of a maximum of 200: 105, 107, 109, 110, 119, 124, 124, 130, 136, and 139, the range being 34 marks. The average range for all the candidates was 33 marks, the smallest range 12, and the highest 63. This range must be regarded as considerable in view of the fact that the examinations were of an elementary character, that the examiners were experienced in this type of work, and that they were marking according to carefully drawn-up marking-schemes.

33. In the type of examination where there are many assistant-examiners the Chief Examiner criticises their marks, and makes adjustments for different standards of marking. The distributions of marks are also sometimes reduced to a standard, but no such adjustment would alter the order of the candidates in the batch assigned to a single assistant-examiner. At a competitive examination of this kind the absolute mark does not matter, as it does in the case of a School Certificate examination. It is only the order that matters, and we must therefore consider this point.

34. The following are the most important results with regard to the first 50 candidates :—

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

Thus 33 candidates would get into the first fifty places whichever couple of examiners marked their scripts; but the fate of the other candidates for the next 17 places would depend on the chance of being assigned to particular couples, the chance of success being greater for some candidates than for others.

35. There is much less agreement with regard to the lowest third of the whole group, so that the element of chance in the award of special places on the plan adopted is very considerable.

36. We now consider Arithmetic and English separately, taking first Arithmetic.

Out of the 150 candidates in Arithmetic there are 63 who got 80 or more marks from at least one examiner, and of these 18 got 80 or more from all examiners. Supposing we regard 80 as a high mark intended to indicate scholarship level, we find complete agreement among the examiners in regard to only 18 out of the 63 possible.

37. The Arithmetic Paper was divided into two parts, Part A and Part B. Part A consisted entirely of twenty straightforward calculations. The variations in dealing with this part were very small, and mainly due to the illegibility of the writing of certain candidates. The average range, *i.e.*, difference between the highest and lowest marks, was only 2·1 per cent. of the maximum, whereas the average range for the two Parts was 14·7 per cent.

38. In spite of the elaborate precautions taken in the marking-scheme, there were very great differences between the examiners in dealing with Part B, which included problems. In a question of which the maximum was 15 marks, one candidate received 15 from one examiner, 12 from three examiners, 8 from two, 7 from two, and 4 from two examiners. But for other candidates there was greater agreement. For 20 candidates the marks were exactly the same, and for 33 the marks only differed by 3 or 4 out of the 15 maximum.

39. The English Paper consisted of two Parts, A and B, of which A was an Essay Paper. A detailed scheme was used for marking the Essay, marks being awarded for the following seven separate elements:—Vocabulary, Accuracy, Craftsmanship, Consistency, Completeness, Substance and Quality. The maximum

for each element was 7 marks.[1] In respect of Vocabulary, only one-third of the candidates got the same mark from as many as five out of the ten examiners.

40. The averages for the different examiners varied considerably. The variation of the averages of the different examiners for the several elements is shown in the following Table :—

	Maximum = 7 marks for each element		
	Highest	*Lowest*	*Range*
Vocabulary	5·93	3·09	2·84
Accuracy	5·36	3·05	2·31
Craftsmanship	4·69	3·20	1·49
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

41. The mean deviations of marks also differed considerably, and the average of the mean deviation of examiners varied from element to element as shown below :—

Vocabulary	*Accuracy*	*Craftsmanship*	*Consistency*	*Completeness*	*Substance*	*Quality*
1·23	1·26	1·25	1·19	1·17	1·22	1·33

42. The paper on English, Part B, dealt mainly with the sense of passages, the sense of phrases, and the sense of single words. Except with regard to one question, for which 66 candidates received the same mark from all the 10 examiners, the agreement was small.

An elaborate analysis has been made of the marks awarded for parts of a question, which cannot easily be summarised here.

Some examiners marked consistently higher, some consistently lower, than the majority; others marked sometimes high, sometimes low, and it is obvious that an examiner who does this will alter the order of the candidates considerably from the order of the majority.

[1] The investigators employed this scheme because it had been used in similar examinations, but are not in any way committed to the view that it is satisfactory.

Special Place Examination (II): *English Essay.*

43. The question-paper gave a choice of one out of four subjects, and the time allowed for the work was 30 minutes.

The main object of the investigation was to compare the results of marking when such essays are marked on impression only, with the results when they are marked in accordance with a detailed marking scheme.

44. Typed copies were made of 15 trial scripts, and circulated to the ten examiners concerned, together with a draft detailed marking scheme. The marking scheme was then amended to meet the criticisms of the examiners, and answers to all doubtful points were furnished to them.

45. Typed copies were then made of 150 other scripts, of which the marks originally allotted to them by the examining authority showed that they varied in marking from very poor to very good. 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, which he could mark for handwriting. The following are the most important instructions 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 (iii) below, 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.

(iii) The maximum mark for all scripts is 100. The examiners were supplied with the amended marking scheme from which the following paragraph is extracted :—

Marks are to be allotted as follows :—

(i) Quantity, quality and control of ideas	50 marks
(ii) Vocabulary	15 marks
(iii) Grammar and Punctuation	15 marks
(iv) Structure of Sentences	10 marks
(v) Spelling	5 marks
(vi) Handwriting	5 marks
Total	100 marks

46. In order to test whether the scripts of Set 1, comprising Nos. 1–75, and those of Set 2, comprising Nos. 76–150, were

approximately equivalent, the sets were re-shuffled and re-numbered, and were then marked by three examiners, X, Y and Z, other than those who took part in the examination of the final scripts. X, Y and Z were all members of the same panel of examiners for a Special Place examination, though not for this particular one. Subsequently the marks allotted by X, Y, and Z were re-grouped according to the original numbers, 1–75, 76–150, and it was found that the average mark allotted by each examiner was the same for Set 1 as for Set 2, although this average differed from examiner to examiner. It was also found that the distribution of the marks of each of the examiners was approximately the same for Set 1 as for Set 2.

It would therefore appear that any difference in the main investigation between the markings of the two Sets by an individual examiner must be due to the difference of method employed, and not to a difference between the two Sets.

47. The first and most striking results of the main investigation are given below :—

AVERAGE MARKS AWARDED BY THE EXAMINERS

	Examiners[1]										*Difference between highest & lowest averages*
	A	B	C	E	G	K	L	M	N	P	
Set 1— (Impression Marking) -	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	

Thus in every case the average mark awarded to Set 2 for scripts marked by details was greater than the average of marks awarded to Set 1 for scripts, marked by impression.

[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 the 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.

With Examiner E the difference between the averages is 27 marks; with Examiners A, B, G, M, P, the difference is about 10 marks, and with only 3 examiners is the difference small. Thus the marking by details produces higher marks on the average than the marking by impression. It is also noteworthy that the averages of the several examiners are closer together when the marking is made by detail than when it is made by impression only. The mean deviation of the averages of the impression marks is 5·2, and that of the averages of the detailed marking only 3·4. Again, the average range of marks was 36·5 for the marking by impression, and 28·9 for the marking by the detailed scheme. The analysis shows that the marking by means of a detailed scheme yields on the whole closer results from the different examiners than the marking by impression.

48. Marking by impression shows very great differences between the examiners. The greatest difference was shown in the marks of a candidate who received the following marks:—50, 63, 69, 15, 78, 62, 75, 48, 71, 64, showing a range of 63. The lowest range was 13, and the average was 36·5. In the marking by details the highest range was 52, in the case of a candidate who received marks varying from 26 to 78, and the lowest range was 14·5. The average range was 28·9.

A detailed analysis of the figures showed that the greater ranges yielded by the marking by impression are not due to a higher figure for 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.

49. The last point is important. It means that the use of a detailed marking scheme does conduce to a closer approximation of the standards of examiners, but that it does nothing to reduce the element of random marking.

50. The difference between the different examiners is very great. In the marking by impression Examiner E awards 50 marks of less than 40, and Examiner C only 2. On the other hand, Examiner L gives 12 marks of 72 or more, and Examiner E gives none. In the marking by details Examiner M gives 21 marks of less than 40, and Examiner C gives none. Examiner E gives 21 marks of 72 or more, and Examiner M gives only 6. There are only two examiners whose marks show approximately

similar distributions, and whose averages are approximately the same when marking by the two different methods.

51. On the other hand, the averages of the two standard deviations for the two methods of marking are 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 different candidates—the same spread of the marks.

52. Although this is true of the averages, some examiners show a different standard deviation in their marks by the two methods.

53. The average ranges for the different elements are shown below.

	Ideas	*Vocabulary*	*Grammar*	*Structure*	*Spelling*	*Handwriting*
Maximum	50	15	15	10	5	5
Average range	19·9	5·5	8·1	4·9	2·1	1·5
Percentage of Maximum	39	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.

54. It will be seen that the greatest average range occurs in grammar, and the least in handwriting. 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.

55. The number of cases (out of the total of 75) in which six or more of the examiners agree are as follows:—Ideas, 39; Vocabulary, 44; Grammar, 20; Structure, 28; Spelling, 48; Handwriting, 63.

56. It has been seen that examiners give higher marks when marking by details than when marking by impression. An attempt was made to discover how the examiners distributed among the various categories of candidates the excess of marks resulting from the second method of marking. It showed that they tended on the whole to favour scripts that were "average" to "just above the average," and to undermark the other categories, especially the "very good"; but the differences were small.

College Entrance Scholarship Examination: English Essay

57. The Paper, which was set at an Entrance Scholarship examination for a group of colleges in a University, gave a choice of four subjects for the essay, but no further instructions. The time allowed was 3 hours.

58. Fifty scripts were selected from a larger number, so as to include those of five holders of scholarships or exhibitions. They comprised the scripts of all the 10 candidates who had selected the first subject; of all the 8 who had selected the second; of all the 11 who had selected the third, and of 21 who had selected the fourth.

59. The examiners were asked to assign numerical marks with a maximum of 100, and also 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.

60. The numerical marks varied considerably. The range of the marks allotted to candidates varied from 7 to 36, and the average range is 19·6 per cent. The extreme cases are shown below:—

Candidate	*Examiners*					*Range*
	A	B	C	D	E	
		Marks Awarded				
No. 25	60	32	65	50	68	36
No. 1	45	38	20	55	20	35
No. 40	40	44	70	75	50	35

The averages of the marks awarded by the different examiners, on the other hand, were close together. They are as follows: A, 51·9; B, 52·7; C, 54·8; D, 54·0; E, 50·6.

61. The following Table shows the statistical distribution of classes by the various examiners:—

CLASSES AWARDED BY THE VARIOUS EXAMINERS

Examiner	*1st Class*	*2nd Class*	*3rd Class*	*4th Class*
A	7	24	15	4
B	8	23	14	5
C	5	29	15	1
D	2	34	12	2
E	5	26	14	5

62. The following Table shows the awards of all the examiners to the 25 candidates who were allotted either a First Class or Fourth Class by any examiner :—

Examiner	A	B	C	D	E
No. of Candidate					
*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.

63. It is especially interesting to see the different selections of candidates by the different examiners for a First Class.

Examiner	A	B	C	D	E
		Candidates	*Awarded a*	*First Class*	
Nos.	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	—	—	—

It will be seen that not a single candidate out of the seventeen was placed in the First Class by more than three out of the five examiners. Three candidates each received three votes; four candidates each received 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.

64. It is noteworthy that though there is comparatively little difference between the averages of the different examiners, the order in which they place the candidates differs greatly. It is quite clear that in an examination of this kind the marks obtained by a candidate are to a very great extent a matter of chance, depending on the particular examiner by whom the essay is marked.

University Mathematical Honours

65. The Paper contained 12 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. Three hours were allowed.

66. Twenty-three scripts were marked independently by six examiners, A, B, C, D, E and F. The scripts were then independently revised by the pairs of examiners AB, CD, EF. There were thus produced six sets of original marks and three sets of revised marks. The nine sets of marks are printed below:—

Maximum Mark = 300

Examiner	A	B	C	D	E	F	*Range*	(A, B)	(C, D)	(E, F)	*Range*
Candidate											
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	3
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[1]	48	55	53	52	47	46		52	52	48	

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

67. It will be seen that the maximum difference of the averages of the individual examiners is about 11 marks—just under 4 per cent. of the maximum mark. The maximum difference of the averages of the three pairs of examiners is 8·5 marks.

68. It is interesting to note that the spread of marks, as measured by the mean deviation, is roughly the same in the case of each examiner and of each pair. There is thus no evidence here that when pairs of examiners allot marks they necessarily award marks with a smaller spread than when they act individually.

69. The differences of the averages yield very little indication of the differences of the marks allotted to individual candidates. The six independent markings of Examiners A to F yield ranges of which the lowest is 17 and the highest 64, with an average of 34·7 on a maximum of 300.

70. The procedure of settling marks on the verdict of two examiners, though it affects the averages very little, had a much greater effect in reducing the ranges, of which the extremes were 3 and 46, and the average 18·3. 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 was assigned 132, 123 and 169 marks, or of Candidate No. 4, who was assigned 186, 177 and 210 marks, is remarkable.

71. It should be noted that the examiners agree in their placing of the first two candidates at the top of the group and in placing the 13th in order of merit. They do not agree in the placing of the other 20. On the other hand, it is noteworthy that the pairing of the examiners notably diminished the difference in the order in which the candidates are placed.

72. The following instances of the difference of opinion between the various examiners are striking:—Candidate No. 1, whose place varies with the individual examiners from 4th (Examiner E) to 12th (Examiner B) of the 23 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). Candidate No. 4 is placed 12th by the pair AB (marks 186), he is placed 12th by the pair CD (marks 177), but is placed 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 in merit.

University History Honours

73. The examination papers were four in number, on the subjects shown below :—

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).

Instead of numerical marking, a scheme of literal marking was adopted in accordance with the practice of most History examinations in this country. Owing to this fact the section on this subject does not lend itself to condensation and is therefore given in full in Appendix I, pp. 59-77, below.

Viva Voce (Interview) Examination

74. 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 at interviews for the selection of candidates for public and private appointments generally.

It appeared, therefore, desirable to test the degree of concurrence of two Boards of Examiners appointed to conduct an examination of this kind.

75. 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 kind of 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.

76. It was decided to offer a prize of £100 on the results of a *viva voce* examination of this kind. The examination was limited to students who were studying, or who had recently studied, at a university, and were certified by the university authorities to be suitable, in their judgment, as candidates for the Junior Grade of the administrative class, Home Civil Service [this is the technical name for the appointments of the highest grade in the Home Civil Service, open to competition]; and the candidates were required to be within the age-limits prescribed for that examination for the year 1934 (21 to 23 on August 1, 1933).

77. The scope of the examination was defined, as in Civil Service regulations (*see* para. 81 (*d*) below).

78. Thirty candidates applied, and of these 16—12 men and 4 women—with excellent University records, were selected for the purpose of the examination. They had received their training in one or more of the following Universities and Colleges:—Oxford, Cambridge, London, Bristol, Glasgow, University College, Nottingham, and University College, Southampton. Each candidate filled in a form similar to that required by the Civil Service examining authorities, to which was attached a confidential report from a tutor or other university authority and a report by the candidate himself on his life and education. Copies of these documents were furnished to each of the examiners.

79. Two Boards were constituted from the following persons:—

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.

80. It was originally intended that the two Boards should have the same number of members, but one of the prospective examiners, the Head of an important college, was accidentally prevented from attending on the morning of the examination, and could not be replaced at the last moment. The examination was held on 27 March, 1934.

81. The following are the more important instructions given to the examiners :—

(*a*) There will be two Boards of Examiners—Board I and Board II. 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.

(*b*) 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.

(*c*) Each candidate is to be examined *for not less than a quarter of an hour and not more than half an hour.*

(*d*) 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.

(*e*) 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

[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 for the sake of simplicity.

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 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.

(*f*) Suitable mark-sheets will be provided.

(*g*) Examiners are requested to sign their mark-sheets and give them in to the Chairman of the Board.

82. At the end of the day, each Board carefully reviewed its marks, in order that the members might be sure that the marks allotted translated correctly their impressions of the relative abilities of the candidates.

83. The marks awarded are set out below :—

MAXIMUM MARK 300

No. of Candidate	*BOARD I* — *Initial mark awarded by examiners before discussion*					*Final mark awarded by Board I*	*BOARD II* — *Initial mark awarded by 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	193
16	180	100	160	180	240	170	220	200	150	190	175

84. The order in which the candidates were placed is shown below :—

Candidate	*Board I Marks*	*Board II Marks*	*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	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	10
10	210	235	8½	6
11	210	236	8½	5
12	230	232	*4	7½
13	120	177	15½	14
14	210	247	8½	4
15	220	193	6	12
16	170	175	13	15½

85. 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.

The prize was awarded to Candidate No. 4, who was placed second by Board II and bracketed fourth by Board I.

86. There were no cases of complete agreement; the closest were the cases of Candidates Nos. 9, 12, and 16 with 10, 2, and 5 marks difference respectively. On the other hand there were extreme cases of disagreement, Candidates Nos. 1, 2, 6, and 7 with 92, 70, 70, and 70 marks difference. The average difference is 37 marks. These 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, and indicate the great influence that this test might have in the final placing of a candidate in a Civil Service examination.

87. The coefficient of correlation between the marks of the two Boards is 0·41. This is comparatively small, and in view of the

*The 3 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.

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.

88. It is probable that the different questions asked of the candidates leading to the different subjects discussed at the two interviews affect the marks finally awarded to the candidates. That the circumstances of the two interviews were entirely different is apparent when we look at the individual assessments of the examiners.

89. In the cases of candidates numbered 13, 3, 1, 2, and 7, the two Boards' marks are entirely different, there is no overlapping. The members of each Board were in agreement within different limits as to the merits of these candidates, and in the case of Candidate No. 1, for instance, the limits are absolutely separated. Board I assessed the merits of this candidate at 120, the individual examiners having awarded marks between 100 and 150; Board II assessed the candidate at 212, the individuals having given marks between 190 and 240.

90. These results show definitely that the evidence on which the examiners could judge the candidate was different in the two cases, that is, that the two interviews were so differently conducted that we might almost suppose different candidates to have been examined. 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.

91. 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.

92. The individual examiners' assessments show very close agreement in certain cases, Board I agreeing within 10 marks in the case of Candidate No. 2, within 30 marks in the case of No. 3, Board II within 30 marks in the case of Candidates Nos. 3, 4, 11, 13, 15.

Some of the marks are widely different. The different examiners of Board I gave to Candidate No. 16 100, 160, 180, 180, and 240 marks; they gave to Candidate No. 4 170, 210, 220, 240

and 280 marks; the examiners of Board II gave to Candidate No. 9 165, 220, 250, and 270 marks.

93. 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.

94. 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 of the marks of individual examiners with the final marks of the Board.

BOARD I

A	B	C	D	E
91	·90	·63	·89	·84

BOARD II

F	G	H	I
·73	·86	·82	·72

95. We find that the evidence shows that 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 pointed out above, the evidence placed before the two Boards was materially different, owing to the inherent nature of an interview of this kind.[1]

[1] I think that my impressions as an impartial and silent observer of the proceedings of the two Boards (having also 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. It was, of course, impossible for me to hear all the candidates examined by both Boards. 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.

Part II—Differences of Standard and Random Variations of Different Examiners

96. In Part I of the investigation the marks allocated to the work of a number of candidates by a number of examiners at different kinds of examinations have been presented and analysed up to a certain point.

Before proceeding further with the analysis, it is desirable to consider briefly the processes by which the marks are obtained.

97. For this purpose it will be convenient to use the phrase "a unit piece of work" to mean any written answer or script which is accorded a mark independently of any other mark accorded to any other piece of work. Thus the phrase may refer to a whole English essay if the essay is marked purely by impression; or it may refer to an answer to a simple arithmetical computation which forms part of a larger question; or it may refer to an element in an answer, such as "Vocabulary" in an essay, if this is accorded a mark separately from other marks accorded to other elements present in the essay.

98. 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 before him the original passage dictated, and in Arithmetic, he would have the answer to a simple sum. In other cases, such a model piece of work may not be available; but the examiner may have clearly defined instructions as to how many marks to allot to a certain answer, how many to take off for a certain type of mistake, and so on. At other times, again, he may have neither a model nor precise instructions to follow, but he will have in his own mind some sort of ideal answer.

99. The possibilities of different marks being accorded to a unit piece of work by a number of examiners are *a priori* obvious. Even when a perfect model exists to which reference may be made, differences of handwriting may give rise to discrepancies; what is illegible to one examiner may be legible to another. When the perfect model does not exist, different examiners may read different meanings into the words and phrases and symbols written in the answer, and so award different marks. Even when the model answer is before the examiner, if it consists of a fairly

lengthy collection of words, examiners may differ in their judgment of what is "like" and what is "unlike" the model. Again the state of health of an examiner may have an effect on his marking as time goes on; his standards of what is perfection may alter, and his judgment may wobble.

100. It is sometimes assumed that if two examiners allot the same average marks, and especially if they allot the same distribution of marks, to a group of scripts, their markings will be identical throughout. Such resemblances may however co-exist with a substantial difference in the marks awarded to individual candidates; for differences of the kind to which we have referred may be present, but may cancel out when averages are taken. Thus, the average of two examiners, and their distribution of marks may be the same, but nevertheless the order in which they place candidates may be different.

101. A practical illustration of the differences of examiners' marks is taken from the investigation on the Special Place Examination, English Paper B. The following are the detailed and the total marks awarded by two examiners, B and D, to the first ten candidates on the roll in this examination, for four questions.

Question	1		2		3		4		*Total*		*Order of Merit*	
Maximum Marks	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 illustration the averages of the total marks are the same, and the averages for the different questions are practically the same; yet in only one case are the marks exactly the same, and in one case they differ by 7. The orders of merit are different.

102. Taking the evidence afforded by this series of scripts marked by the two examiners, we might fairly judge that they

both marked the four questions in this paper according to the same standards; but the individual idiosyncracies of the two examiners are shown in the marks awarded to the candidates in respect of the various questions, and are not entirely eliminated from the totals, which therefore exhibit discrepancies.

103. It may happen that in addition to the kind of discrepancies noted in the foregoing illustration one examiner may on the average tend to award higher marks than another examiner for each unit piece of work, so that his average mark for a whole script will be higher. This kind of difference between two examiners will always be revealed by an examination of average marks, but it may accompany discrepancies of the kind already referred to.

104. The assumptions made and conventions used in this part of the analysis are as follows :—

(*a*) That a piece of work is worth a definite number of marks in a scale.

(*b*) That this mark would be allotted by the "perfect examiner." We call this mark the "ideal" mark.

(*c*) That every examiner attempts to discover this ideal mark but may fail (i) because his standard of marking differs from the ideal, and (ii) because he introduces random variations into his marking.[1]

(*d*) That an examiner who introduces a large random element into his marking is not as precise an examiner as one who introduces a small random element into his marking.

(*e*) That a first approximation to the ideal mark may be obtained by taking the simple average mark of a number of examiners; and that a closer approximation may be obtained, if we take account of the fact that some examiners are more precise than others, and if we therefore use a "weighted" average, the "weight" of an examiner being inversely proportional to the variance of his random variations.

105. These assumptions make it possible for us to split up any group of marks awarded by examiners to a number of scripts into the following components for each script :—

[1] As a further refinement, there is the possibility that the individual examiner may differ from the "perfect" examiner also in the "spreading" of the ideal marks. When examiners are used to team work and are supplied with detailed instructions as to marking, including instructions relating to the standards of Pass, Fail, Credit, Honours, etc., and are accustomed to the kind of examining work which they have undertaken, it might be argued that there is not much likelihood of differences of this nature being introduced into the results of the marking. But, in order to test the extent of this kind of discrepancy, three sets of our data were submitted to a new analysis based on this assumption. It was found that the general conclusions relating to the incidence of the random element in the marking still held good.

(i) The ideal mark; (ii) the amount by which the examiner's standard differs from the ideal; (iii) the random element appropriate to that particular script.

For an outline of the method by which the ideal marks have been calculated, see para. 133 below.

106. The size of the random element is estimated by means of the standard deviation[1] of the group of random variations present in the marks allotted by an examiner, and this measure can therefore be used to compare one examiner with another as to precision of marking, an examiner with a large standard deviation being considered as less precise in his marking than one with a smaller standard deviation. We can also compare one paper in a subject with another paper, or one subject with another from the point of view of precision of marking by observing the differences between these standard deviations.

School Certificate History

107. 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 on School Certificate History.

Candidate's Number.	*"Ideal."*	*EXAMINER B.*			*EXAMINER H.*		
		Constant Difference	*Random Variations*	*Original Marks.*	*Constant Difference*	*Random Variations*	*Original Marks.*
1	42	−9·8	+0·8	33	+6·3	−0·3	48
2	38	−9·8	−0·2	28	+6·3	+1·7	46
3	44	−9·8	−3·2	31	+6·3	−1·3	49
4	48	−9·8	+1·8	40	+6·3	−2·3	52
5	43	−9·8	+0·8	34	+6·3	−2·3	47
6	47	−9·8	−0·2	37	+6·3	−2·3	51
7	52	−9·8	+5·8	48	+6·3	−0·3	58
8	38	−9·8	−5·2	23	+6·3	−3·3	41
9	36	−9·8	+1·8	28	+6·3	+3·7	46
10	46	−9·8	+4·8	41	+6·3	+2·7	55
11	44	−9·8	−4·2	30	+6·3	−1·3	49
12	55	−9·8	−5·2	40	+6·3	−0·3	61
13	39	−9·8	−3·2	26	+6·3	−1·3	44
14	43	−9·8	+7·8	41	+6·3	+3·7	53
15	56	−9·8	−2·2	44	+6·3	+2·7	65
	Average		Standard Deviation	Average		Standard Deviation	Average
	44·7	−9·8	3·9	34·9	+6·3	2·3	51·0

[1] The standard deviation of a series of numbers is the square root of the average of the squares of the differences of the numbers from their average.

The random element is larger with Examiner B than with Examiner H, and this is reflected in the higher standard deviation of his random variations.

108. The standard deviations indicating the extent of the random element in marking at the two investigations on School Certificate History are given below.

Examiner	*Standard Deviation of Random Variations*		*Order of Examiners*[1] *according to size of Standard Deviation (Smallest S.D. placed first)*	
	1st Investigation	*2nd Investigation*	*1st Investigation*	*2nd Investigation*
A	4·9	—	—	—
B	3·9	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 are very similar. 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 correlations 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.

109. 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 (say) 2·3, would

[1] Examiner A is omitted from this Table as he did not take part in the second investigation.

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 examination this range of marks would include the border line marks for Pass and for Credit. Thus a candidate who is possibly worthy of a Credit may actually achieve only a Pass or even be dubbed a Failure, or he may succeed in being given a mark of Credit instead of a Pass.

110. The extent of the variability amongst the candidates, due to their differences 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 difference in the standards of marking of the examiners combined with the random variations which, in view of the sizes of the standard deviations, are likely to occur, 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 Body. As stated in para. 5 above, the scripts all received the same "middling" mark.

School Certificate Latin.

111. The Table below shews the standard deviations of the random variations of the two groups of examiners in Latin.

GROUP 1.		*GROUP* 2.	
Examiner A	1·45	*Examiner* 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	Average	2·40
Standard Deviation of Ideal Marks	3·76	Standard Deviation of Ideal Marks	3·65

112. This investigation gives material from which a comparison is possible of the precision of marking for the two parts of the Paper. Some examiners appear to be relatively more precise when marking Paper I (grammar, etc.) than Paper II (prescribed books), but with others the contrary is the case, and it is doubtful if the evidence warrants the drawing of a general conclusion either way. The standard deviations are shown below.

PAPER I. (*Maximum 50 marks.*)				*PAPER II.* (*Maximum 50 marks.*)			
Group 1.		*Group* 2.		*Group* 1.		*Group* 2.	
A	0·98	G	1·48	A	1·77	G	2·20
B	0·70	H	0·93	B	1·01	H	1·35
C	1·68	J	2·32	C	1·50	J	1·49
D	1·77	K	1·20	D	1·37	K	1·74
E	1·02	L	1·78	E	1·31	L	1·19
F	0·62	M	1·08	F	1·08	M	0·85
		N	1·81			N	2·49
Average	1·13		1·51		1·34		1·62

Of the examiners of Group 1, A, B, E and F mark Paper I with more precision than Paper II; of the examiners of Group 2, G, H, K and N mark Paper I with more precision than Paper II; in one case four out of the six examiners, in the other four out of the seven mark Paper I with more precision than Paper II.

School Certificate French

113. 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.[1]

114. 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 below, together with the classified results :—

	BOARD I.					*BOARD II.*				
Candidate.	*Ideal*	*Exr. C.*	*Difference between Exr. C. and Ideal*	*Awards*		*Ideal*	*Exr. J.*	*Difference between Exr. J. and Ideal*	*Awards*	
				Ideal	*Exr. C.*				*Ideal*	*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	52	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

In this Table D=Distinction; C=Credit; P=Pass; F=Fail.

[1]The detailed instructions to examiners naturally lead to precision of marking.

	BOARD I.					BOARD II.				
Candidate.	Ideal	Exr. C.	Difference between Exr. C. and Ideal	Awards		Ideal	Exr. J.	Difference between Exr. J. and Ideal	Awards	
				Ideal	Exr. C.				Ideal	Exr. J.
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	C
47	54	53	−1	C	C	57	54	−3	C	D
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

In this Table D=Distinction; C=Credit; P=Pass; F=Fail.

115. The candidates whose class is affected by the random element in Examiner C's marking are Nos. 4, 14, 15, 16, 21, 27, 39, 41, eight in all. The details are as follows :—

Candidate	Difference between C.'s mark and Ideal	Ideal Class	C.'s Class	Raised + Lowered −
4	+4	P	C	+
14	−1	P	F	−
15	−2	C	P	−
16	−3	P	F	−
21	+6	C	D	+
27	−9	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 in class in 5 cases.

Similarly the candidates whose class is affected by the random element in Examiner J's marking are Nos. 15, 18, 19, 25, 30, 35, 39, seven in all. The details are as follows :—

Candidate	*Difference between J.'s mark and Ideal*	*Ideal Class*	*J.'s Class*	*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 in class in 4 cases.

These two illustrations are typical of the effect of the random element on the class 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 8 candidates, and in the other case 7 candidates out of 50, have their class altered owing to the presence of the random element in the examiner's marks.

116. An examination of the standard deviations of the examiners' random variations obtained when individual questions in the papers are the subject of consideration reveals the fact that some questions lead to more precise marking on the part of the examiners than others. For instance, answers to Qn. 1 of Paper I receive more precise marking than answers to Qn. 2 of Paper II.

School Certificate Chemistry

117. The standard deviations of the random marks are shown below :—

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	Standard Deviation of Ideal Marks	19·8

The random element is not very pronounced, ranging from about 2½ to 5½ 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.

118. One of the chief reasons why the members of the two Boards placed different numbers of candidates in the various grades, Distinction, Credit, Pass, Fail, is that the two Boards on this occasion adopted different borderline marks for these grades.

School Certificate English

119. The random variations introduced into the marking are indicated below (with a maximum mark = 100):—

Examiner	A	B	C	D	E	F	G
Standard Deviation	4·12	4·56	3·27	3·84	3·12	3·00	4·27

120. The marks awarded by the examiners to the seven questions in this examination which were answered by the majority of the candidates were submitted to the same method of analysis with the results given below, where for comparative purposes each figure has been reduced to a percentage of the maximum marks per question.

STANDARD DEVIATIONS

	Essay	*Précis*	II 1	II 4	II 5	II 10	II 13
Ideal Marks	7·3	8·5	11·7	10·4	12·4	10·6	13·6
Examiners							
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

121. 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.

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.

Paper I deals with Essay and Précis; and the marking of this Paper is less precise than the marking of Paper II, which deals mainly with set-books.

The total variation of the candidates' marking may be regarded as due to a combination of their natural variation with the variation of the examiners' marks. Where the variation of the examiners is comparatively large, as in the marking of Paper I, the total variation is mainly due to the variation of the examiner. Where it is smaller, as in the case of Paper II, the total variation is mainly due to the natural variation.

Special Place Examination (II): *English Essay*

122. Our method of analysis enables us to give a reasonably clear answer to the question:—"Is marking by details more precise than marking by impression?" We saw that the detailed marking gave on the whole higher average marks than marking by impression, but the random element appears to be present to the same degree in both types of marking.

123. The table below gives the standard deviations of the random variations:—

Examiner	*Marking by Impression*	*Marking by Details*
A	10·0	7·7
B	9·0	11·0
C	9·0	7·9
E	9·8	10·0
G	11·5	6·0
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	8·4	7·9

Five examiners (A, C, G, M, P) have less of the random element in their marking by details than in that by impression, while the other five are more precise when marking by impression than when marking by details. On the average there seems no ground for asserting that either method of marking is better than the other from the point of view of precision.

College Entrance Scholarship: *English Essay*

124. Whereas the differences between the examiners' average marks are rather small, and consequently the standards of marking of the examiners are on the whole very little different from the ideal, the random variations are, on the other hand, rather large.

Standard deviations:—A, 6·8; B, 9·1; C, 9·0; D, 7·5; E, 6·5.

The large discrepancies between the different examiners in this investigation are due more to the random element in the marking than to any steady differences of standard.

125. The data of this investigation were further analysed with the object of discovering what influence, if any, the subject of the essay had on the resultant mark. There were four essay subjects, and the analysis showed that there were considerable differences between the marks awarded by different examiners to essays on different subjects. Thus the average of Examiner A's marks for the candidates who wrote on Subject No. 2 was 9 marks (out of 100) more than that of Examiner D, but A's average for Subject No. 4 was 5½ less than D's average for that subject.

126. The fate of a candidate in this type of examination is partly dependent on the particular examiner's reaction to the subject of the essay.

University Mathematical Honours.

127. The standard deviations of the random variations in the marking are reduced when the examiners are grouped in pairs for the revision of the marks. The table below shows the standard deviations, based on a maximum of 100.

Examiner	A	B	C	D	E	F
Standard Deviation	4·2	3·8	4·0	3·5	4·1	4·3

Pair of Examiners	G	H	J
Standard Deviation	2·3	3·2	3·0

128. The differences between the different examiners' standards of marking are not very great, and these were reduced when the revision took place; but the differences of standard still remaining, coupled with the random element, would still have the effect that in certain cases the class awarded to a candidate would depend on the pair of examiners by whom he was marked.

University History Honours

129. The method which was used to estimate the size of the random element in the marking in the previous investigations is no longer possible of application in the present case as the marks are given in literal form. But by a modification of the method used we can get the relationship between the standard deviation of the random variations and the ideal marks for each examiner.

130. We find comparatively large random variations present in the marks allotted by the examiners in this investigation, the standard deviations being in many cases greater than the standard deviation of the ideal marks. As there was a general consensus of opinion amongst the examiners that the candidates were on the whole below the first class, we can assume that the standard deviation of the ideal marks of each paper would be 10 out of 100 on a numerical basis. On this assumption, the corresponding average standard deviations of the random variations for the four papers of the examination would be :—Paper I, 12 ; Paper II, 17 ; Paper III, 10 ; Paper IV, 9 marks (out of 100). Not too much precision should be accorded to these figures ; they are mainly estimated with the idea of comparing the results of this investigation with the others where the marks were given numerically and not literally.

Summary of the foregoing Sections.

131. The following table gives average figures for standard deviations of the random variations, two figures being given where two Boards or two Groups of Examiners acted separately. In each case the marks are referred to a maximum of 100.

School Certificate, Latin	1·9 ; 2·4
French	2·9 ; 2·7
English	3·6
Chemistry	3·5 ; 3·7
Honours Mathematics	4·0 ; (pairs, 2·9)
School Certificate History	5·2 ; 5·6
English Scholarship Essay	7·7
English Essay Special Place Examination	8·4 ; 7·9
Honours History	12 ; 17 ; 10 ; 9 (Four papers)

132. As might have been expected, the most precise results are obtained in those examinations where many detailed instructions are given, and where the marking is therefore standardised as much as possible, and the least precision is obtained in the examinations of the essay type, where far more is left to the judgment of the examiner.

Method of Calculating Ideal Marks.

133. Let us call the marks awarded to the pieces of work written by n candidates by the several examiners X_t, Y_t, Z_t, ... , where t takes all values from 1 to n. We assume that the "ideal" mark appropriate to the piece of work of the t'th candidate is Q_t, and that $X_t=Q_t+A_t$, $Y_t=Q_t+B_t$, $Z_t=Q_t+C_t$, and so on,[1] A, B, C, ... , being used to indicate the differences between the ideal marks and those awarded by the various examiners A, B, C, ...

The averages of these various marks for the group of n candidates are indicated by $\overline{X}$, $\overline{Y}$, $\overline{Z}$, ... , $\overline{Q}$, $\overline{A}$, $\overline{B}$, $\overline{C}$, ...

Deviations from the averages are indicated by small letters x_t, y_t, z_t, ... , q_t, a_t, b_t, c_t, ...

Then we have $x_t=q_t+a_t$, $y_t=q_t+b_t$, and similarly.

Consider the pair $x_t=q_t+a_t$, $y_t=q_t+b_t$. We have

$$x_t-y_t=a_t-b_t,$$

and $(x_t-y_t)^2=a_t^2+b_t^2-2\,a_t\,b_t$.

Summing such identities for $t=1$ to n, gives

$$S(a_t^2)+S(b_t^2)=S(x_t-y_t)^2$$

assuming that $S(a_t\,b_t)=0$, an assumption which depends on the random element in A's marking being independent of the random element in B's marking.

---

[1] The further refinement referred to in the footnote to para. 104 (p. 38) would correspond to a modification of this assumption. We should now assume

$X_t=r_a\,Q_t+A_t$, $Y_t=r_b\,Q_t+B_t$, $Z_t=r_c\,Q_t+C_t$,

and so on, the r's being multipliers differing from one examiner to another. The statistical analysis is naturally modified in consequence. This subject is discussed in more detail in two memoranda by Professor Cyril Burt and by Dr. Rhodes in *The Marks of Examiners*.

Similarly we can obtain the equation,

$$S(a_i)^2+S(c_i)^2=S(x_i-z_i)^2,$$

on a similar assumption.

If there are m examiners, there are $\frac{m(m-1)}{2}$ such equations.

From these equations we can estimate the most probable values of each of $S(a_i^2)$, $S(b_i^2)$, ... , because in each of these equations the right hand side is known from the data.

We obtain our results in this form :

$$S(a_i)^2=\frac{m}{m-2}\,S(x_i^2)-\frac{2}{m-2}\,S(x_ip_i)$$

$$-\frac{1}{(m-1)(m-2)}\,[S(x_i^2)+S(y_i^2)+\ldots]+\frac{1}{(m-1)(m-2)}\,S(p_i^2),$$

where $p_i=x_i+y_i+\ldots$.

$$S(b_i^2)=\frac{m}{m-2}\,S(y_i^2)-\frac{2}{m-2}\,S(y_ip_i)$$

$$-\frac{1}{(m-1)(m-2)}\,[S(x_i^2)+S(y_i^2)+\ldots]+\frac{1}{(m-1)(m-2)}\,S(p_i^2).$$

and so on.

These estimates of $S(a_i^2)$, $S(b_i^2)$, ... , being proportional to the variances of the random elements introduced by A, B, ... , into the marking give us weights w_a, w_b, ... , from which the ideal marks may be estimated. Thus

$$Q_i=\frac{w_a X_i+w_b Y_i+\ldots}{w_a+w_b+\ldots}$$

APPENDIX I

UNIVERSITY HISTORY HONOURS (DETAILS OF INVESTIGATION)

1. *Character of 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.

2. *Procedure.*—The University concerned furnished us with all the scripts available in the subjects enumerated above from a recent Honours examination.[1] Unfortunately 3 scripts (which happened to be among the best) had been accidentally destroyed. The total number of scripts available was 18 for Paper I, 17 for Paper II, 18 for Paper III, and 16 for Paper IV.

[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.

The following 17 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.

Professor A. Browning, M.A., D.Litt., 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.Litt., 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.

3. The scripts of Paper I were marked by 5 examiners; the scripts of each of the other papers by 10 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.

4. 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 1

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)		

5. 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.

6. 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.

7. 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 the numbers 1 to 24 to the successive grades, δ to $\alpha+$, and that, side by side with the literal tables, we have inserted numerical tables on this basis. But, for the reasons stated above, the numbers indicating grades must not be regarded as numerical marks. They are ordinal numbers, not cardinal.

8. 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, $\alpha+$, 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.

9. An index of the examiners who marked the various papers is given in the Table below :—

TABLE 2

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.

10. In Tables 3, 4, 4a, 5, 5a, 6 and 6a 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. 7 and 8 above.

TABLE 3.

Paper I.

	Marks allotted					*Numerical representation of the the marks in ordered grades*						
Examiner	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 4

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 4A

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	2	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 5

Paper 111

Marks allotted.

Examiner		A	B	F	H	J	K	L	N	Q	R
Cand. No.	1	βα	β?+	β	α=	β+	β+	β?+	β	γβ	β+
	2	β−	β+	α=	β+	β?+	β	β?+	β−	β+?+	β+
	3	β?+	β++	β	β++	β	β+?+	β+	βγ	βγ	α
	4	β?−	β	β	β	β?+	β	β?−	βγ	γ	β?+
	5	βα	γ+	β−	β−	β+	β+	β?−	γβ	γ+	β−
	6	β++	α−	βα	β+	β++	β?+	β+?+	β?+	γβ	αβ
	7	β	β	β−	β	β−	β=	β	β=	βγ	β+
	8	β	α=	β	β	β+	β−	β+	β=	βγ	β+
	9	α	γ+	β+	γβ	α−	β+?+	β+	β+	αβ	α−
	10	β?+	β+	β++	β−	γ+	β++	β?−	β	β++	β?+
	11	δ	γ	β=	β	γ−	βγ	γβ	δ	γ−	γ+
	12	β?+	γ+	β+	β	β	β?−	βγ	β=	γ+	β
	13	β+	β−	α−	β+	β+	βα	β?−	β+	β−	β?+
	14	β+?+	βα	αβ	γ+	γ+	β+?+	β?+	β	β+	βα
	15	β	β++	β−	γβ	β	γ+	β=	β	γ+	β+?+
	16	β+	α−	αβ	βα	βα	β?+	βα	β+?+	αβ	βα−
	17	β?+	γ+	β=	β?+	γ	γ	βγ	γ	γ	β?−
	18	β	β++	α=	β++	β++	β	βα	β+	βα	α?+
Median		β?+	β+ / β?+	β?+	β	β?+	β / β?+	β / β?+	β?−	βγ	β+

TABLE 5A

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	15½	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 6

Paper IV

Marks allotted

Examiner	A	B	E	F	G	H	J	L	M	Q
Cand. No. 1	β	β?+	β−	β?−	β?−	β?+	β+?+	β?−	β+?+	βγ
2	β?−	β−	β?+	βα	β+	β+	α=	β?+	β+	β?+
3	β+	β+	β+	β	β+?+	β−	β++	β?+	β?−	β−
4	β+?+	β	βα	β?+	β−?−	β	β+	βγ	β+	βγ
5	β	β?−	β?−	β=	β−	γ+	γ+	γβ	γβ	γ+
6	βα	β+	β?−	β?−	α=	βγ	β++	β?+	β+?+	γβ
7	βγ	γ+	β=	β?−	β?+	βγ	γ+	β=	β	γ
8	βγ	α−	β+	β+	β=	β?+	γ+	β?+	β+	βγ
9	α	β+	βα	β?−	αβ	β?+	β+	β+	βα	β+?+
10	βα	β?+	β?+	αβ	β?−	β+	β?+	β−	β+	β+?+
11	δ	γ	γ+	β?−	β−	γ	γ+	γβ	β=	βγ
12	β+?+	β	β?+	β?+	β?+	βγ	β+	βγ	β	γ+
13	βα	α−	α−?−	α=	β?+	β++	β+?+	β=	αβ	β+
15	β+?+	β+	αβ	β?+	β+	β+	γ+	β−	β++	β−
17	β	β	β−	β−?−	β+	β=	β+	β=	β+	γ+
18	β?−	β++	αβ	βα	αβ	α−	β?+	β+?+	β+?+	αβ
Median	β?+	β?+	β?+	β / β?+	β?+	β / β?+	β+	β−	β+	βγ

TABLE 6A

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

11. 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.

12. 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. 4), 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 $\alpha\dagger+$, 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. $\beta\dagger-$ 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 $\beta-\dagger-$ 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.

13. 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 border line 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.

14. We give in Tables 7 to 10 below the classification statistics of the various examiners on the foregoing basis, for the scripts marked by them.

TABLE 7

PAPER I (Ancient & Mediæval History)

Mark	*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)	β+ (13)	β (11)	β− (9)	γβ (5)

Thus Examiner D gives two candidates clear α's, 4 candidates a borderline mark between α and β, 10 candidates β, 1 candidate γβ, and 1 candidate γ. Q returns them all as β or worse, and no examiner uses δ.

15. TABLE 8

PAPER II (Mediæval and Modern History)

Mark	*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	β+ (13)	β+?+ (14)	β?+ (12)	β?+ (12)	β?− (10)	β− (9)	β+ (13)	β?− (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.

16. TABLE 9

PAPER III (Essay)

Mark	*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)	β+ / β?+ (12-13)	β?+ (12)	β (11)	β?+ (12)	β / β?+ (11-12)	β / β?+ (11-12)	β?− (10)	βγ (6)	β+ (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.

17. TABLE 10

PAPER IV (Political Theory)

Mark	*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)	β− (9)	β+ (13)	βγ (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.

18. 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.)

19. 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*	*III*	*IV*
B	β+?+ (14)	β+ (13) β?+ (12)	β?+ (12)
L	β?− (10)	β (11) β?+ (12)	β− (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) β+ (13)	(B) β+?+ (14)	(R) β+ (13)	(J & M) β+ (13)
Lowest	(P) β− (9)	(J) β− (9)	(N) β?− (10)	(L) β− (9)
Difference (Number of grades)	4	5	3	4

20. 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.

21. 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 3 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 4 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 5 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 6 shows that Candidate No. 8 gets $\alpha-$ from B and $\gamma+$ from J, a range of 16 grades.

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.

22. The discrepancies between the marks awarded by the examiners which have been the subject of discussion in the preceding paragraphs may be considered to be due to two causes (1) constant differences of standard of marking on the part of examiners (2) the presence of an element of randomness in an examiner's marking.

These points are discussed in Part II above (see p. 42 *et seq.*).

APPENDIX II.

BRIEF SUMMARY OF THE WORK OF THE FRENCH INTERNATIONAL INSTITUTE EXAMINATIONS ENQUIRY

(Commission Française pour l'Enquête Carnegie, sur les examens et concours en France).

1. The French Committee[1], who have received every assistance from the French Ministry of Public Instruction, have published a general report on French examinations, their character, the spirit by which they are inspired, and their relationship to the national system of education in the form of an *Atlas de l'enseignement en France* (in-quarto-raisin, pp. xiii, 183, 13 planches hors texte, à Paris, à la Maison du Livre, 4 Rue Félibien, 75 francs).

2. They also issued a questionnaire to some 4,000 persons with regard to certain examinations, and will publish a summary of the replies.

3. They have carried out a series of investigations on the *baccalauréat* examination, in many ways similar to the investigations described in the present pamphlet, and the results have been recorded in a volume entitled *La correction des épreuves écrites dans les examens, enquête expérimentale sur le baccalauréat* (in-quarto-raisin, à Paris, à la Maison du Livre, 4 Rue Félibien).

4. The first examination investigated by the Committee was the *baccalauréat*, because in their view this examination is both the most typical and the most important of all the French examinations. In the University of Paris alone there are about 15,000 candidates annually for the two parts of the *baccalauréat*. The examination serves both as a school-leaving examination for the lycées (both for boys and girls) and as an entrance examination to universities and to the liberal professions. "It is," says the French Committee, "an instrument of selection of what maybe called the directing classes" (*l'instrument de sélection des classes dites dirigeantes*)[2].

[1]The personnel of the French Committee is given on page 7 above.

[2]It is clear from the context that the phrase "directing classes" is used here to designate not classes privileged by birth but those who actually exercise a directing influence in the social system. The phrase was used in the same sense in the Report of the Auxiliary Committee on Education of the Indian Statutory Commission (1929).

5. The two parts of the French *baccalauréat* correspond, roughly speaking, to the examinations for the School Certificate and for the Higher School Certificate in England. The first part is normally taken at the age of about 16 by pupils of the *classe de première* (formerly called the *classe de rhétorique*). The second part is normally taken a year later by pupils in two parallel classes, the *classe de philosophie* and the *classe de mathématiques*. In these classes philosophy is treated as the most important subject on the literary side, mathematics as the most important on the scientific side ; but mathematics and other science subjects are taught in the *classe de philosophie*, while philosophy and other literary subjects are taught in the *classe de mathématiques*.

6. Both parts of the *baccalauréat* include a written examination and a *viva voce* examination in a number of subjects. Only those who pass on the written examination are admitted to the *viva voce*. A total aggregate of 50 per cent on the subjects of the written examination is required for a candidate to be admissible to the *viva voce* examination—it would appear, without a minimum requirement in any one subject.

7. The following summary is translated from the proofs of Chapter VIII of the volume :—

(1) Two investigations have been undertaken by the French Committee (Commission Française Carnegie) on the marking of scripts at the *baccalauréat* examination. The chief investigation was undertaken with reference to the examinations in :

Translation from Latin (*Version latine*)	Part I of the *baccalauréat*
French Essay (*Composition française*)	
English	
Mathematics	
Philosophy	Part II of the *baccalauréat* for pupils of the *classe de philosophie*
Physics	Part II of the *baccalauréat* for pupils of the *classe de mathématiques*

100 scripts corresponding to each of these examinations, which had been actually written at the examinations held in July, 1930, were corrected and marked by 5 examiners

(*correcteurs*) chosen from the panel of examiners for the *baccalauréat* (the actual mark of the examiner at the *baccalauréat* examination furnishing a sixth mark).

The scripts chosen formed a sufficiently typical sample of the *baccalauréat* scripts as a whole.

A supplementary investigation was made on three French essays (*copies de composition française*), selected from those used for the principal investigation, which were corrected and marked by 76 different examiners.

(2) The maximum ranges[1] of the marks attributed to one and the same script in the first investigation by the different examiners were as follows :—

12 marks out of 20 for Latin translation	[60 per cent.]
13 marks out of 20 for French Essay	[65 per cent.]
9 marks out of 20 for English	[45 per cent.]
9 marks out of 20 for Mathematics	[45 per cent.]
12 marks out of 20 for Philosophy	[60 per cent.]
8 marks out of 20 for Physics	[40 per cent.]

The mean differences between the marks of two examiners varied from 1·88 out of 20 in Physics [*i.e.*, 9·40 per cent.] to 3·36 out of 20 in Philosophy [*i.e.*, 16·80 per cent.].[2] The number of the differences between two examiners equal to or higher than 5 marks out of 20 (25 per cent.) was 2·5 per cent. in Physics and 23 per cent. in Philosophy.

(3) The number of scripts which were recorded as deserving an average mark or a mark higher than the average in the opinion of some of the examiners (but not of all) was as follows :—

Latin translation	... 50 per cent. of the total number of the scripts
French Essay	70 per cent. of the total number of the scripts
English	... 47 per cent. of the total number of the scripts
Mathematics	36 per cent. of the total number of the scripts
Philosophy	... 81 per cent. of the total number of the scripts
Physics	... 50 per cent. of the total number of the scripts

[1] The term 'range' is used, as in the text of this pamphlet, to denote the difference between the highest and lowest marks allotted by different examiners to the same script.

[2] The differences between each pair of examiners for each candidate were calculated, there being, with six examiners, 15 differences in respect of each candidate, and 1,500 for each subject.

8. For the second investigation on the French Essay three scripts, Nos. 23, 25 and 34, were selected, each of which at the original *baccalauréat* examination had been awarded 36 marks out of 80 (or 45 per cent.) and had been ranked as 24th out of a batch of 50. These three scripts were marked independently by 76 examiners. The marks for script No. 23 varied from 4 to 52, for script No. 25 from 12 to 64, and for script No. 34 from 16 to 56 out of a maximum of 80. The mean marks for the three scripts were as follows: Script No. 25—25·9; Script No. 25—40·0; Script No. 34—34·4.

9. The book contains an elaborate statistical analysis of the relations between the marks of the different examiners, from which the following may be quoted :—

> After reduction by means of appropriate corrections of the scales of the different examiners to the same level of severity (by reducing to the same average) and to the same distribution (by altering the marks so that they have the same standard deviation), there still remain important differences between the results of the pairs of examiners. The correlation between the marks of two examiners was never perfect, with a value of $r = 1$, and was as low as $r = 0{\cdot}112$ (correlation between the marks of Examiner C and Examiner D in Philosophy for 50 scripts of women candidates). The mean correlation coefficient of all the examiners taken in pairs varies from $r = 0{\cdot}429$ in Philosophy (scripts of women candidates) to $r = 0{\cdot}888$ in Mathematics (scripts of male candidates).

Waterlow & Sons Limited, Printers, London and Dunstable.