Schools Council Examinations Bulletin 29

Comparability of standards between subjects

Evans | Methuen Educational

Nuttall, D.L., Backhouse, J.K., & Willmott, A.S. (1974). *Comparability of standards between subjects*. Schools Council Examinations Bulletin 29. London: Evans/Methuen.

SCHOOLS COUNCIL EXAMINATIONS BULLETIN 29

Comparability of standards between subjects

D. L. Nuttall, J. K. Backhouse
and A. S. Willmott

Evans/Methuen Educational

First published 1974 for the Schools Council
by Evans Brothers Limited
Montague House, Russell Square, London WC1B 5BX
and Methuen Educational Limited
11 New Fetter Lane, London EC4P 4EE

Distributed in the US by Citation Press
Scholastic Magazines Inc., 50 West 44th Street
New York, NY 10036
and in Canada by Scholastic–TAB Publications Ltd
123 Newkirk Road
Richmond Hill, Ontario

ISBN 0 423 89130 8

Printed in Great Britain by
Richard Clay (The Chaucer Press) Ltd,
Bungay, Suffolk

Contents

Foreword

In its early years the Schools Council gave priority in examinations research to monitoring the standards of the new CSE examination. After studying four consecutive years of the CSE in operation, the Council commissioned the National Foundation for Educational Research to conduct a new series of research studies aimed at providing evidence which could inform judgements on which future examination policy and practice might be based. This report is one of the results of the research.

The report considers an aspect of comparability which has long been of interest and importance to pupils, teachers, examiners and 'users', although, as the report indicates, concern has tended to increase in recent years. Stated briefly, the question is whether or not an examination result in any one subject is 'comparable' to the same result in another subject. The stress often laid by schools and users on the total number of subjects in which a candidate gains certification might be taken to imply such comparability. This report provides evidence which throws some doubt on such an implication. The procedures from which this evidence is derived are of particular interest for two main reasons.

First, the maintenance of educational standards is rightly attracting a great deal of public interest, and on this subject there is perhaps a natural tendency for some people to make generalizations and value judgements based on limited evidence. The fact is that comparability presents researchers with technical and procedural problems of great complexity, so that it is difficult to produce firm, irrefutable evidence. This report describes in detail the main procedures which were adopted to investigate comparability between subjects—procedures which, moreover, make a contribution to the methodology of comparability studies generally. Some of the technical details which will be of interest primarily to research workers are presented in the appendices. The report itself has been presented in a descriptive form which, it is hoped, will be meaningful to a much wider audience.

Secondly, the report highlights an issue on which opinion is divided. The procedures adopted are based exclusively on statistical evidence which is *external* to that of individual subjects. It can be argued that this is reasonable in that examination results ought to reflect the general ability of candidates, thus ensuring comparability. Equally, however, it can be argued that what distinguishes one subject from another are the characteristics peculiar to that

subject, in which case external factors have little or no relevance. Behind this difference of view there is a more immediate problem. Traditionally when examiners mark candidates' work they base their assessments on the evidence they have before them and candidates are rewarded for the level of attainment they have reached. If the suggestions advanced tentatively in the last chapter of this report were implemented this would no longer be the case, as examiners would be subjected to an external constraint in order that comparability would be assured. The problem, moreover, might be most acute in those subjects in which, on the evidence of this report, there might be marked differences between the performance of boys and girls.

The report is published as a contribution to the general discussion about the meaning of grade standards; its publication does not imply acceptance or otherwise by the Schools Council of the concept of subject comparability nor the adoption of any view by the Council on the suggestions made in the report.

Abstract

The investigations reported here stem from the examinations research programme conducted by the Research into CSE and the Research into GCE projects at the National Foundation for Educational Research under the sponsorship of the Schools Council.

The report consists of a description of methodological studies in the area of comparability of standards in different subjects. Except in the case of two boards, the data were collected in the 1968 CSE Monitoring Experiment and the results must be treated with caution since they relate to the examinations of summer 1968.

Two main methods were employed: one used external reference tests and the other internal evidence, namely the examination grades achieved in other subjects. The methods were found to lead to essentially the same results. It was concluded that, with the samples used and treating the sexes together in both the GCE O-level sector and the CSE sector, English (language and literature) and possibly art appeared to be consistently leniently graded and that chemistry and French appeared to be consistently severely graded. Further, in the GCE sector, physics appeared to be severely graded and, in the CSE sector, mathematics appeared to be severely graded.

The question of sex differences in examination performance was also investigated. The pattern of differences between subjects was not the same for boys and girls analysed separately. The performance of boys was worse in French and English literature than expected on the basis of their overall performance and better than expected in mathematics and geography, while the performance of girls was worse than expected in chemistry and physics and better than expected in English language and English literature.

The necessary assumptions of the methods are critically discussed. Accepting the validity of these assumptions, the implications of the findings for grading procedures and comparability of standards are explored. The report concludes with a series of questions raised by the results of the research about the nature of examination grading and standards, and it is hoped that they will stimulate public discussion of these controversial topics.

Acknowledgements

The authors would like to thank all those who have helped to make this report possible and regret that space does not permit the mention of the names of all of them. Our special thanks must go to Pat Evison and Malcolm Killcross (both have since left the NFER), who contributed much to the development of our thinking and who spent many hours collating data and compiling tables. We would also like to thank our fellow researchers, both inside and outside the NFER, who have helped us to clarify our ideas; in particular, we wish to thank Gerry Forrest of the Joint Matriculation Board. More recently, the high standard of typing of Jenny Walker and Dianne Horton has made the final stages of the work much less of a burden than might have been the case and our thanks are due to them both, particularly to Dianne who had the unenviable task of typing the final draft in double-quick time.

We acknowledge with thanks the permission of the Joint Matriculation Board to reproduce results from their Occasional Publications 33 and 34, of the South-Western Examinations Board to reproduce an extract from their Seventh Annual Report, and the kindness of one CSE board in supplying us with the 1971 examination results of all their Mode I candidates. We also thank the staffs of the NFER Statistical Services, the Oxford University Computing Service and the Atlas Computer Laboratory.

Last, but not least, we should like to thank the Schools Council for its continuing sponsorship of research into examining at 16+ at the NFER.

1. Introduction

The problem of comparability

The comparability of standards in public examinations has long been a topic of debate among teachers, parents and those whose job it is to select applicants for employment or for further and higher education. Myths and misconceptions abound: winter examinations are easier than summer examinations; board A's English examination is easier than board B's; CSE grade 1 is easier (or more difficult) to obtain than a corresponding pass at GCE O level. These myths often arise as a result of teachers' experience in individual schools; late each summer one can often find letters in the educational press from heads who find, perhaps, that of the thirty candidates they entered for board A's examinations, twenty-five passed, but that when the same candidates took board B's examinations, only ten passed. Need there be concern about this sort of occurrence? There are a large number of factors that might have given rise to it, none of which has anything to do with lack of comparability of standards between the two boards. The candidates might have concentrated far more on the syllabus of board A than on that of board B; their teacher might have been more inspired by the content and approach of board A's syllabus, or the candidates might have tried harder with board A's examination because it was the first they took.

This example illustrates the fact that the comparability of standards in examinations is by no means an easy topic to study. It is relatively rare for candidates to take the examinations of two different boards in the same subject; it is educationally undesirable in any case and the results are of dubious value. Rather, comparability has to be studied in terms of probabilities and likelihood. This is best illustrated by the definition of CSE grade 1, as given in Examinations Bulletin No. 1: 'a 16-year-old pupil whose ability is such that he might reasonably have secured a pass in the O level of the GCE examination, had he applied himself to a course of study leading to that examination, may reasonably expect to secure grade 1, having followed a course of study [regarded by teachers of the subject as appropriate to his age, ability and aptitude]'.[1]

This does *not* mean that if a candidate obtained a CSE grade 1 he would have got a pass in O level if he had taken it. It means that had he followed the O-level course instead of the CSE course, there is a high probability that he would have obtained a pass.

Examination boards have always been at great pains to ensure comparability of standards between themselves, but until the establishment of the CSE there

was little published research about comparability. In 1964, the Schools Council commissioned the National Foundation for Educational Research to conduct annual investigations from 1965 to 1968 into the comparability of standards among the fourteen CSE boards, and between the fourteen CSE boards and the eight GCE boards as far as CSE grade 1 was concerned. Four reports were published and the final one, reviewing all the studies, concluded that, apart from a very few isolated exceptions, there was no evidence to suggest any lack of comparability in standards between boards in any of the six major subjects studied (English, mathematics, history, geography, French and science).[2,3,4,5]

The main method adopted in all these studies is particularly relevant to much of the work reported in later chapters and demonstrates the necessarily indirect, probabilistic approach to comparability. Representative samples of candidates in each board were studied and, because of different entry patterns between regions (probably caused mainly by regional differences in the proportion of children who leave school at 15), the average grade in English obtained by these samples of candidates differs region by region. But this does not necessarily imply any lack of comparability. What is required is some objective measure of the differences in the nature of the candidate entry in each region. In these studies, the objective measure used as a common yardstick to compare regions was a test of scholastic aptitude. Every candidate in the sample from each region took this test, and it was found that the average test scores varied from region to region. Could the differences between the average grades in English be attributed just to the differences between the average scholastic aptitude of the candidates? To find out, statistical procedures are required and these are described in greater detail in Chapter II and Appendix A; in essence, these procedures are used to predict what the average grade in English for candidates in a region *would have been*, given their average scholastic aptitude test score, if all boards were applying the same standards. The predicted average grade is then compared with the actual average grade achieved by the sample of candidates; any differences between predicted and actual grades are attributed to a lack of comparability of standards.

The most important assumption made in using these procedures is that the scholastic aptitude test can be used to predict performance in an English examination or, in other words, that there is a demonstrable relationship between scholastic aptitude and English attainment. The research shows this to be the case, but the relationship is modest. An alternative approach was therefore adopted in the 1966 and 1968 studies. An objective test of English attainment was employed in place of the scholastic aptitude test. The relationship between the candidates' performance on the attainment test and their performance in the examination was much stronger than that between the apti-

tude test and the examination. The test of English attainment is also intuitively more appealing than the aptitude test because of its greater relevance and similarity to the examination itself. Nevertheless, the two types of test lead to the same results time and time again.[3,5,6] The key assumption would thus appear to be convincingly validated (in the context of public examinations at 16+).

An alternative approach, again with an intuitive appeal greater than that of the aptitude test method, is cross-moderation. In this procedure, examiners from different boards independently grade a sample of scripts. If each board is applying the same standards, then the average grade awarded by each examiner should be the same and this has indeed proved largely to be the case (see, for example, the South Western Examinations Board's *Bristol Experiment*,[7] and *CSE: an Enquiry into Standards in Four Subjects by Four Boards*.[8] In the only comparison between the aptitude test method and cross-moderation both methods gave rise to the same result.[3] The major assumption in cross-moderation exercises is that examiners will be able to apply their own board's standards to examination scripts based on syllabuses rather different from their own in an experimental situation.

A fairly long account has been given of the different methods employed to study the comparability of standards between examining boards to illustrate the fact that all the methods depend upon a number of assumptions which may be challenged. The findings thus allow a number of interpretations of which a lack of comparability is only one, but probably the most likely (see Schools Council Working Paper 21,[3] p. 4). But comparability between examining boards is only one of the facets of comparability, albeit the one that has attracted the most publicity and the most research. The following five aspects of comparability may be identified:

a between examining boards in the same subject and year;
b between years in the same subject within an examination board;
c between modes of examining in the same subject and year within an examination board;
d between alternative syllabuses in the same subject and year, which may perhaps be considered a special case of **c**;
e between subjects in the same year within an examination board.

These five aspects of comparability are all currently being investigated by the Research into 16+ Examinations Project at the NFER under the sponsorship of the Schools Council, with particular emphasis on **b** and **c**. This report concentrates on **e**, comparability between subjects, but throws up a sixth aspect, comparability between the sexes, which is discussed in Chapter IX. The aim of this report is to compare different methods of investigating comparability

between subjects, and to air some of the methodological and educational problems of subject comparability.

What is subject comparability?

It is easy to understand what comparability between examining boards means and why there is a need to be sure that it has been achieved. In a national examinations system, there would be obvious unfairness to individual children if an employer assumed that a CSE grade 3, in a given subject, meant the same throughout the country when in fact this was not the case. It is, however, less easy to understand what subject comparability means.

We (the three authors) argue as follows: we do not expect an individual candidate to achieve the same grade in every subject that he takes. However, we can see no logical reason why, if a large *group* of candidates representative of the population took, for example, both English and mathematics, their average grades should not be the same. Since these two subjects are crucial in gaining further qualifications or employment, there is no reason to suppose that the candidates as a whole would not try equally hard in both subjects. There is also no reason to suppose, again for the group as a whole, that their teachers in one subject are better than their teachers in the other. Although some individuals in the group will be better at English than they are at mathematics, there is no reason to suppose that these individuals will not be balanced by another group who are better at mathematics than they are at English. We therefore argue that their mean grades should be the same.

We obtained from one CSE board the results of *all* the 7019 candidates who took both English and mathematics in summer 1971. Their average grade in English was 2·96, and their average grade in mathematics was 3·55. We submit that a major cause, if not the whole cause, of this difference of over half a grade is a basic lack of comparability between the grading standards used in the two subjects. We have heard it argued that this difference simply reflects the relative difficulty of the two subjects – English is inherently easier than mathematics – but this is to miss the point of the grading system, at least in the CSE sector. Grade 4 is defined in Examinations Bulletin No. 1 as follows: 'a 16-year-old pupil of average ability who has applied himself to a course of study regarded by teachers of the subject as appropriate to his age, ability and aptitude, may reasonably expect to secure grade 4'.[1] Standards are therefore related, subject by subject, to the notional attainment of the average 16-year-old. The fact that one subject is harder than another subject, even if this is indeed the case, has no bearing on grade 4 at all.

The next question is whether differences between standards in subjects matter, in the same way that differences between the standards of examining

boards would matter. We again submit that they do. We offer one simple example of the way in which they might matter: if an employer were faced with two boys, each with five O-level passes, he would have no basis on which to choose between them in terms of their attainment. If, however, both had passes in English language and mathematics, and one had passes in English literature, biology and French while the other had passes in history, physics and Latin, it is highly probably that the latter's general level of achievement would be higher than the former's, on the basis of the results presented later in this report.

A review of previous research

We are not, of course, breaking new ground in suggesting that subject comparability should be investigated nor in proposing ways of doing it. It would appear, however, that interest in the topic is relatively new (except perhaps in the sphere of university final examinations, where the variation in the proportion of Firsts awarded by different faculties often provokes discussion!).

One example, quoted by Brentini, is the case of pure and applied mathematics at A-level.[9] He reports that, in one board, those candidates taking both applied and pure mathematics were achieving lower grades on average in applied mathematics than in pure mathematics. As a consequence, the board has changed the standards in applied mathematics so that those candidates who take both subjects now receive the same average grades in both subjects.

Three CSE boards have also been concerned about the problem. In its *Seventh Annual Report*, the South Western Examinations Board raised its doubts about standards in English and mathematics:

> The main results table shows that there was again a marked difference between the English and mathematics results and the Board's Examination Committee therefore ordered an investigation to be undertaken. This took the form of an enquiry into the patterns of entry in the two subjects and was conducted in the autumn term following the examinations. The results of the enquiry showed that in 1971 the number of pupils taught English and mathematics in the fifth year of all participating schools was 21 773 and 21 081 respectively and that:
>
> i fewer of them were entered for O-level mathematics (35·8 per cent) than English (46·9 per cent)
> ii more were double-entered for CSE and O level in mathematics (14·1 per cent) than in English (9·2 per cent)
> iii more were withheld from entry for both examinations in mathematics (17·1 per cent) than English (5·2 per cent).

> Each of those comparisons appeared to justify more CSE grade 1 candidates and fewer grade 5 and ungraded candidates in mathematics than in English. But in the 1971 examinations the figures were:
>
> grade 1: mathematics 14·9 per cent, English 18·1 per cent
> grade 5 and ungraded: mathematics 30·4 per cent, English 12·5 per cent.
>
> The results of the enquiry were circulated to all schools and it was requested that joint meetings of English and mathematics teachers should be held in every school and later in every advisory group to attempt to identify the reasons for the disparity between the two subjects.[10]

The Board's *Eighth Annual Report* notes that no significant conclusions emerged from any of these meetings.[11]

The Yorkshire Regional Examinations Board has also investigated subject comparability, but on a much broader front,[12] and is continuing its investigations at the present time. Following the work of Forrest,[13] which is discussed in more detail below, the West Yorkshire and Lindsey Regional Examining Board is using a scholastic aptitude test to investigate subject comparability, inter alia, in co-operation with the NFER. Elsewhere, the Australian State of Victoria has recently investigated subject comparability and found some marked differences between subjects. As a consequence it has adjusted standards such that every subject has the same standard.[14] Similar findings have also been made in New Zealand.[15]

By far the most significant study so far reported is that of Forrest carried out within the Joint Matriculation Board.[13] He tested representative samples of 1970 candidates with the NFER's Aptitude Test 100 (see Schools Council Working Paper 34[5]) in each of the following subjects: art, biology, chemistry, English language (with separate samples for two alternative papers), English literature, French, geography, history, mathematics and physics. He found that the average Aptitude Test 100 score of candidates in physics and chemistry was significantly higher than the average test score of candidates in English language and English literature. In itself that finding is not very surprising; the most likely explanation is simply that some subjects are more selective in their entry than others: that is, while the majority of O-level candidates take English language and literature, only the more able take physics and chemistry. Nuttall detected a similar effect in the CSE sector as a whole,[5] and his results are shown in Table I.1. Candidates for physics, chemistry and French, for example, have average test scores about half a standard deviation higher than those of candidates in English and religious education. (See also a small study by Wort.)[16]

The really surprising fact that emerged from Forrest's results was that, generally speaking, the higher the mean test score for any subject, the worse the

Table I.1 Mean Aptitude Test 100 score and number of CSE candidates by subject[a]

Subject	*Number of candidates*	*Mean test score*
Biology	1346	37·4
Chemistry	925	42·8
Domestic studies	1117	32·7
English	5534	36·1
French	1971	43·7
General science	804	37·1
Geography	3313	37·4
History	2630	37·6
Mathematics	5872	39·3
Metalwork	987	37·0
Needlework	865	32·8
Physics	1849	40·8
Religious education	1124	34·5
Woodwork	980	37·1

[a] From Tables 2, 6, 14 and 16 in Schools Council Working Paper 34.[5] The standard deviation of the test scores in each subject is of the order of 10 points and the maximum possible score is 80 points.

mean grade in that subject. In other words, subjects which attracted candidates of higher aptitude on the average tended to be those which awarded the worse grades on average. From his results he concluded that physics and chemistry were noticeably severe in their grade awards, while English language (Paper B) and English literature were noticeably lenient with respect to the average of all subjects under scrutiny. The results in any one year are, however, not totally convincing; if the same picture were to emerge over a number of years, with different samples of candidates, the evidence would be incontrovertible. The study has been repeated with samples of 1971 candidates by Forrest and Smith,[17] and the findings are very similar to those of 1970. A recent study by the Welsh GCE board has also produced similar findings.[18]

Incontrovertibility of evidence does not imply incontrovertibility of interpretation of the evidence. Instead of differences between standards in different subjects, Forrest acknowledges the possibility of bias in the test as a cause of the observed differences (i.e. the nature of the test content may be such that candidates taking mathematics, for example, may do better simply by virtue of their studies than their peers who are not taking mathematics). This report offers evidence on this issue in Chapter II. Forrest also acknowledged the possibility

of differences in test and examination performance between boys and girls being a factor which would complicate the interpretation of the results and this issue is discussed in Chapter IX of this report.

Past and current work does, therefore, suggest that there may be a lack of comparability of standards between different subjects, but not enough research has been done and not enough discussion of the topic has taken place to discover how serious a problem it is, if a problem it be. The aim of this report, as already indicated, is to look at possible methods of investigating the problem and to raise issues for public debate. We stress that the results we present are not of significance in themselves; for reasons explained below they must be interpreted with great care and in any case they relate to the situation as it was in 1968 and not necessarily to the situation as it is now.

The samples used

Except where indicated, the data used throughout this report are those collected in the 1968 CSE Monitoring Experiment. They therefore relate to samples of candidates who were tested with the scholastic Aptitude Test 100 in February or March 1968 and who sat CSE or O level (or both) in summer 1968. Details of the sampling procedure and, in the case of CSE boards, the size of samples, mean test scores, mean grades and correlations between the test scores and the grades awarded are given in Schools Council Working Paper 34.[5] Similar details about the GCE board samples are given in Appendix B. While the CSE samples have been shown to be representative of the population of CSE candidates region by region, there is no claim that the samples of GCE candidates are representative of their respective populations (i.e. 16-year-old school pupils taking O level in summer 1968). There is no published evidence which allows a check on their representativeness to be made. The results themselves must therefore be treated with extreme caution; the fact that English language in our board 1 sample appeared to be 0·73 of a grade lenient *cannot* be generalized to conclude that board 1 was 0·73 too lenient on *all* its candidates in English language in summer 1968.

Outline of the report

Chapters II to V describe four different methods of investigating subject comparability. Each one uses data from the same GCE board (2) to exemplify the method. Chapter II concentrates on the use of tests to investigate standards in different subjects and explores the issue of possible bias in the tests, while Chapters III, IV and V employ methods which do not require the use of a test but rely in essence on a comparison of the performance of the same candidates

in different subjects. In all four cases the methods are presented as simply as possible, and technical details of the more complex methods may be found in Appendix A. Chapter VI compares the results produced by these different methods, and Chapter VII describes the results of applying the methods to the examination grades of the complete population of candidates in one CSE board and investigates sampling procedures. Chapter VIII examines the results themselves and discusses the significance of the consistency of the patterns of subject differences that occur. Chapter IX discusses sex differences in examination performance, and Chapter X summarizes the conclusions of the investigations and discusses the problems that would arise if any adjustments of standards were to be made.

References

1. Secondary School Examinations Council. *The Certificate of Secondary Education: Some Suggestions for Teachers and Examiners* (Examinations Bulletin No. 1). HMSO, 1963.
2. Schools Council. *The 1965 CSE Monitoring Experiment* (Working Paper No. 6, Parts I and II). HMSO, 1966.
3. LARRY S. SKURNIK and JOHN HALL, *The 1966 CSE Monitoring Experiment* (Schools Council Working Paper No. 21). HMSO, 1969.
4. LARRY S. SKURNIK and IAN CONNAUGHTON, *The 1967 CSE Monitoring Experiment* (Schools Council Working Paper 30). Evans/Methuen Educational, 1970.
5. D. L. NUTTALL, *The 1968 CSE Monitoring Experiment* (Schools Council Working Paper 34). Evans/Methuen Educational, 1971.
6. LARRY S. SKURNIK, *Monitoring Grade Standards in English* (Schools Council Working Paper 49). Evans/Methuen Educational, 1974.
7. *Certificate of Secondary Education: Bristol Experiment* (Report on an interboard moderation exercise, November 1967). South Western Examinations Board, Bristol, 1968.
8. *CSE: an Enquiry into Standards in Four Subjects by Four Boards.* Associated Lancashire Schools Examining Board, North Regional Examinations Board, North Western Secondary School Examinations Board, Yorkshire Regional Examinations Board, 1968.
9. ERIC BRENTINI, 'Aspects of comparability in GCE', *Secondary Education*, 2 (Spring 1972), 3–6.
10. *Seventh Annual Report.* South Western Examinations Board, Bristol, 1972.
11. *Eighth Annual Report.* South Western Examinations Board, Bristol, 1973.
12. *A Comparison of Awards by Subject Panels for the period 1966–1970* (Research Report 10). Yorkshire Regional Examinations Board, Harrogate, 1970.

13. G. M. FORREST, *Standards in Subjects at the Ordinary Level of the GCE, June 1970* (Occasional Publication 33). Joint Matriculation Board, Manchester, 1971.
14. L. MACKAY and G. WHITTLE, personal communication. 1972.
15. W. B. ELLEY and I. D. LIVINGSTONE, *External Examinations and Internal Assessments*. New Zealand Council for Educational Research, 1972.
16. RICHARD WORT, 'Academic injustice', *Higher Education Review*, **3** (Summer 1971), 57–60.
17. G. M. FORREST and G. A. SMITH, *Standards in Subjects at the Ordinary Level of the GCE, June 1971* (Occasional Publication 34). Joint Matriculation Board, Manchester, 1972.
18. *An Investigation into the Standards of Subjects in the GCE Ordinary Level Examinations* (Research Report No. 1). Welsh Joint Education Committee, Cardiff, 1973.

II. The use of tests

This chapter is concerned with the use of different kinds of tests to investigate comparability of standards between subjects. The main method of collecting and analysing the data is identical in principle to that used in previous comparability exercises (for example, that of Nuttall[1]). In each board, samples of candidates took Aptitude Test 100, and the examining boards supplied the CSE grades and the unofficial O-level grades that these candidates obtained in all the subjects that they sat in summer 1968. The results of the research discussed in this chapter relate only to the following subjects: art, biology, chemistry, English language, English literature, French, geography, history, mathematics and physics (the same ten subjects as were studied by Forrest[2]).

The analyses are performed in terms of average (unofficial) grades rather than with pass percentages, since it is much more satisfactory statistically to work with complete distributions rather than at specific points in the distributions. There is, however, a close relationship between average grades and pass percentages, as Table II.1 reveals: in general the higher the pass percentage, the better the average grades, as is to be expected.

Table II.1 Sample mean grades and pass percentages in GCE boards 2, 3 and 4

	GCE BOARD					
	2		3		4	
Subject	*Mean grade*	*Pass percentage*	*Mean grade*	*Pass percentage*	*Mean grade*	*Pass percentage*
Art	5·27	65·4	4·99	69·0	5·16	76·6
Biology	5·35	62·3	5·26	64·4	5·44	64·6
Chemistry	5·45	61·7	5·26	68·0	5·85	53·9
English language	5·14	69·8	5·24	67·7	5·15	69·2
English literature	5·30	66·0	5·40	67·1	5·69	64·2
French	5·56	59·8	5·85	52·9	5·72	58·1
Geography	5·57	58·3	5·02	68·8	5·26	66·9
History	5·61	58·4	5·67	55·9	5·66	64·4
Mathematics	5·45	61·7	5·00	72·9	5·22	69·7
Physics	5·44	59·9	5·19	65·9	5·52	63·9
All entries	5·41	63·0	5·29	65·8	5·46	65·3

Test 100: the regression method

Figure 1 shows the average Test 100 score for candidates in each of the ten subjects plotted against their average grade in each subject for GCE board 2.

The pattern in this figure demonstrates the same trend as the one detected in the research of Forrest,[2] namely that the groups of candidates with the highest average test scores tend to be those with the worst average grades.

There can be no doubt, therefore, that some sort of difference between subjects exists. The regression method provides one way of estimating the extent of these differences. The starting point is the 'consensus' standard of the ten subjects included in the analysis: this is defined as the average of the average grades in each subject and the average test scores, and in the case of GCE board 2 these averages are 5·41 and 51·5 respectively.† Since the average relationship between the grades and the test scores across all ten subjects is known (see Table B.5, p. 99), it is possible to predict the average grade that would be expected for any given average test score if grades in each subject were awarded using the same standards. (The method is described in more detail in Appendix A, p. 93.)

For example, the mean test score of the chemistry candidates in the sample for GCE board 2 was 55·0 (3·5 points better than the average test score of the complete sample). The regression method predicts that the corresponding mean grade for a mean test score of 55·0 should be 5·11, on the assumption that grades in chemistry were awarded on the same standard as grades in the other nine subjects. The mean grade in chemistry actually awarded was 5·44. Chemistry is therefore identified as being severely graded by 0·33 grades (5·44 — 5·11 = 0·33).

This process is repeated for each subject in turn, and the results are shown in Table II.2.

Because of sampling error, estimates of the degree of severity or leniency can only be considered to call for notice if they exceed half a grade. In Table II.2, only English language and art have values approaching half a grade. There is, in fact, some doubt as to whether art may validly be included in such an analysis, because the correlation between Test 100 scores and art grades is consistently much lower than the correlations between Test 100 scores and the grades in the other nine subjects. Forrest showed, however, that the exclusion of art made very little difference to the estimates of severity and leniency in the other subjects.[2]

† The values for each subject are given in Appendix B, Tables B.3 and B.4.

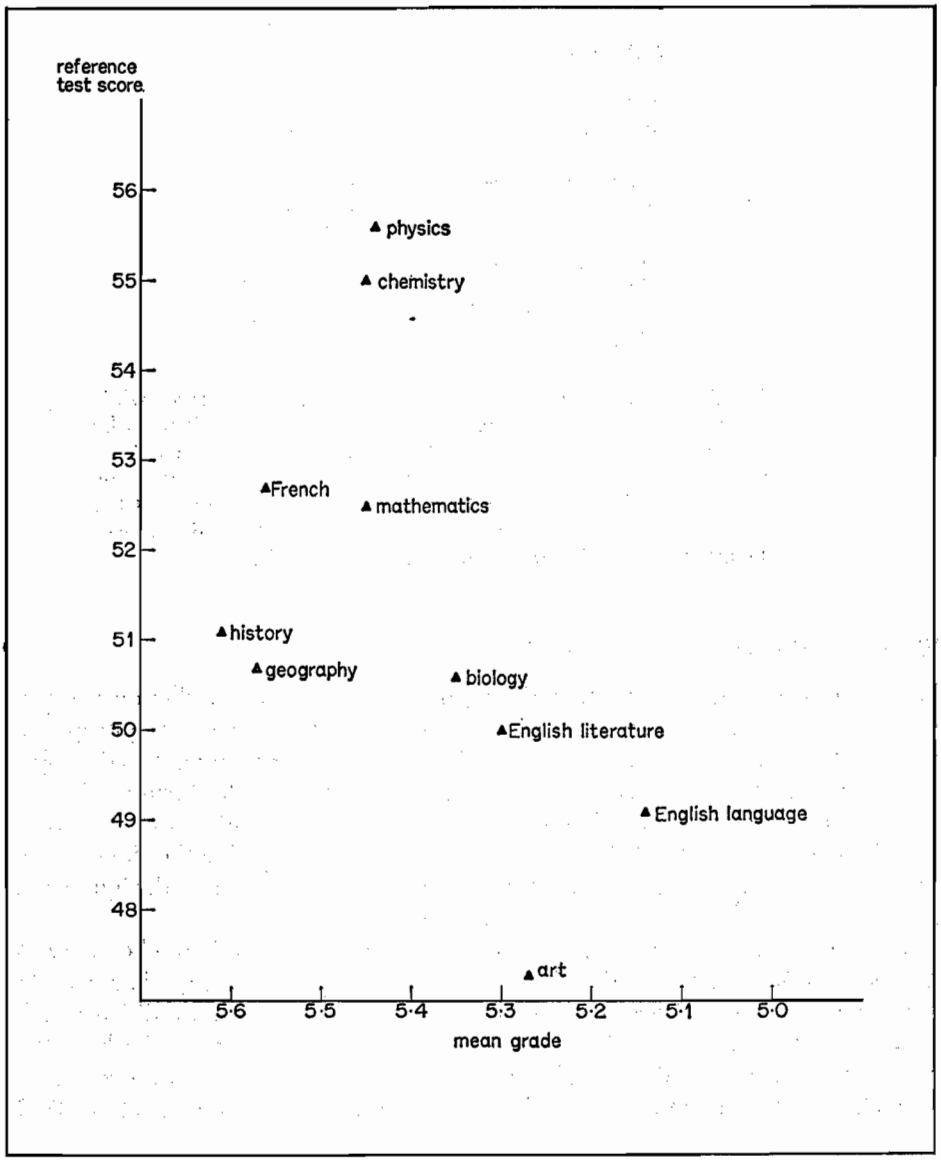

Fig. 1 *Plot of mean reference test score against mean grade for GCE board 2.*

Table II.2 Sample estimates of mean grade severity in GCE board 2 (regression method)[a]

Subject	*Estimate of severity*
Art	−0·49
Biology	−0·14
Chemistry	0·33
English language	−0·49
English literature	−0·24
French	0·25
Geography	0·09
History	0·16
Mathematics	0·12
Physics	0·37

[a] Positive values indicate a tendency towards severity, while negative values indicate a tendency towards leniency. It must be stressed that the terms 'severity' and 'leniency' are a convenient technical shorthand. They do not necessarily indicate that the grading standards in the subject are out of line; only if all the assumptions of the various methods are met can these differences be attributed to differences in grading standards.

Test 100: the guideline method

It has been argued recently that the regression method of predicting average grades for a given test score as outlined above is not the most appropriate method.[3,4] As an alternative, Please and Peaker propose a method called 'structural regression' which uses more information; in the particular form used here, the method employs data relating to the consensus of standards of the CSE boards in addition to the GCE board 2 consensus to determine the slope of the line of prediction. This version of structural regression is known as the 'guideline method' and details are given in Appendix A, p. 93.

As an example of the prediction made by the guideline method, chemistry in GCE board 2 is again considered. The observed sample average grade was 5·44; the regression method predicted an average grade of 5·11 for candidates with an average test score of 55·0 on the assumption that grading standards in the ten subjects were comparable. The guideline method predicts an average grade of 4·82.† The regression method therefore provides a sample estimate of 0·33

† Technically, this is a gross over-simplification, but is justifiable in the context of the argument. For details, see Appendix A, p. 93.

grades severity in board 2 chemistry, while the guideline method provides a sample estimate of 0·62 grades severity.

Table II.3 compares the sample estimates of severity and leniency of the ten

Table II.3 Sample estimates of mean grade severity in GCE boards 2 to 4 (regression and guideline methods)

subjects in the GCE boards 2, 3 and 4 obtained by the regression and guideline

	GCE BOARD					
	2		3		4	
Subject	*Regression*	*Guideline*	*Regression*	*Guideline*	*Regression*	*Guideline*
Art	−0·49	−0·80*	−0·74*	−1·16*	−0·76*	−0·93*
Biology	−0·14	−0·19	0·05	0·17	−0·07	−0·09
Chemistry	0·33	0·62*	0·46	1·01*	0·79*	0·90*
English language	−0·49	−0·57*	−0·33	−0·60*	−0·64*	−0·72*
English literature	−0·24	−0·32	−0·05	−0·26	0·10	−0·08
French	0·25	0·29	0·62*	0·45	0·29	0·37
Geography	0·09	−0·05	−0·42	−0·42	−0·15	−0·02
History	0·16	0·04	0·31	0·04	0·14	0·01
Mathematics	0·12	0·18	−0·29	−0·12	−0·25	−0·15
Physics	0·37	0·72*	0·32	0·85*	0·48	0·75*

methods. Estimates greater than 0·50 grades have been indicated with an asterisk to draw attention to the most important points of comparison.

Since it is the extreme positive and negative values that call for notice, these results indicate a fair degree of consistency as between the regression and guideline methods within each board. The estimates of severity produced by the guideline method exhibit a marked tendency to be greater than those produced by the regression method at the extremes (there are twelve asterisks in the guideline columns but only five in the regression columns).

In the context of comparability of standards between subjects, it is concluded that the guideline and regression methods lead to much the same results despite their different statistical assumptions. (For a discussion of the possible reasons, see Appendix A, pp. 95–6.)

Bias in the test

As mentioned in Chapter I, Forrest noted that the observed differences between subjects might be explained in part, if not in total, in terms of a bias in the test rather than in terms of a lack of comparability of standards between subjects.

The nature of the items in Test 100 is such that it would be reasonable to hypothesize that those candidates entered for mathematics or for science subjects would obtain significantly higher scores on the test than candidates in other subjects simply by virtue of their having followed mathematically oriented courses. In other words, the test scores for the different groups of candidates might not be directly comparable. This hypothesis cannot be tested directly with the data presented in this report, but in the 1968 CSE Monitoring Experiment a test of attainment in English language was used in one GCE board (5) and one CSE board (15). This permits an investigation of test bias from another angle: estimates of severity may be calculated by the regression method using different tests and the results compared.

The test of English attainment was Test E2, which is described fully in Schools Council Working Paper 9.[5] It consists of twenty multiple-choice questions on a passage from D. H. Lawrence, a further twenty multiple-choice sentence-completion items and an inter-linear exercise, worth 20 marks, involving the correction of errors of punctuation, grammar and spelling. Previous work (for example, that of Skurnik and Hall[6]) has shown that performance on Test E2 demonstrates a much stronger relationship to performance in CSE English than does performance on Test 100.† It would be reasonable to hypothesize that Test E2 would not be biased in favour of mathematics or science candidates.

As an additional but possibly less powerful check, the verbal sub-score of Test 100 was used to investigate test bias in GCE board 5 and CSE board 15. Each candidate in the samples for these boards thus had three test scores: Test 100 total score, Test 100 verbal sub-score and E2 total score. Details of sample sizes, mean test scores, mean grades in the ten subjects, and correlations between the test scores and subject grades appear in Tables B.6–9, pp. 99–101.

Table II.4 presents the sample estimates of mean grade severity produced by the regression method employed with each of the three test scores in GCE board 5. Table II.5 presents the same information for CSE board 15. In both boards, the estimates based on the use of each test agree so well that test bias can be ruled out as a factor complicating the interpretation of results of investigations into subject comparability (Forrest and Smith[7] reached the same conclusion). However, the slight differences that exist in the estimates of severity are generally in the direction expected on the hypothesis that Test 100 favours those studying mathematics or science subjects.

† Mean correlation across all boards in 1966 between E2 scores and CSE grades in English = 0·52 (see Schools Council Working Paper No. 21,[6] p. 75). Mean correlation across all boards in 1968 between Test 100 scores and CSE English grades = 0·37 (see Schools Council Working Paper 34,[1] p. 52).

Table II.4 Sample estimates of mean grade severity in GCE Board 5 (regression method)[a]

Subject	TEST SCORE EMPLOYED		
	Test 100 total score	*Test 100 verbal sub-score*	*E2 total score*
Art	−0·82*	−0·67*	−0·54*
Biology	0·56*	0·52*	0·62*
Chemistry	1·04*	1·03*	0·85*
English language	−0·75*	−0·72*	−0·68*
English literature	−0·11	−0·12	−0·02
French	−0·04	−0·08	0·01
Geography	0·20	0·21	0·17
History	−0·38	−0·30	−0·19
Mathematics	−0·46	−0·55*	−0·57*
Physics	0·77*	0·67*	0·36

[a] Asterisks are used to draw attention to values greater than ±0·50 grades; they do not imply the existence of statistically significant differences.

Table II.5 Sample estimates of mean grade severity in CSE board 15 (regression method)[a]

Subject	TEST SCORE EMPLOYED		
	Test 100 total score	*Test 100 verbal sub-score*	*E2 total score*
Art	−0·28	−0·27	−0·35*
Biology	−0·21	−0·19	−0·11
Chemistry	0·23	0·21	0·22
English[b]	−0·27	−0·25	−0·29
French[c]	−0·57*	−0·61*	−0·44*
Geography	−0·01	−0·01	−0·06
History	0·15	0·20	0·14
Mathematics	0·50*	0·47*	0·47*
Physics	0·46*	0·46*	0·41*

[a] The 'tolerance limits' in CSE are ±0·33 grades,[1] so values in excess of ±0·33 have been asterisked. As with GCE board 5, these limits are not statistical confidence limits.

[b] CSE board 15 has a single subject – English – rather than the two subjects – English language and English literature – that most GCE boards provide.

[c] Estimates are based on only 26 pupils and the correlations between the Test 100 scores and grades in French are negative (Table B.9, p. 101), so the results for French should be treated with extreme caution.

References

1. D. L. NUTTALL, *The 1968 CSE Monitoring Experiment* (Schools Council Working Paper 34). Evans/Methuen Educational, 1971.
2. G. M. FORREST, *Standards in Subjects at the Ordinary Level of the GCE, June 1970* (Occasional Publication 33). Joint Matriculation Board, Manchester, 1971.
3. N. W. PLEASE, 'The 1965 CSE Monitoring Experiment: a comment', *Educational Research*, **13** (June 1971), 233–5.
4. G. F. PEAKER, 'The 1965 CSE Monitoring Experiment: a reply [to comment by Please]', *Educational Research*, **13** (June 1971), 235–6.
5. JACK WRIGLEY, F. H. SPARROW and F. C. INGLIS, *Standards in CSE and GCE: English and Mathematics* (Schools Council Working Paper No. 9). HMSO, 1967.
6. LARRY S. SKURNIK and JOHN HALL, *The 1966 CSE Monitoring Experiment* (Schools Council Working Paper No. 21). HMSO, 1969.
7. G. M. FORREST and G. A. SMITH, *Standards in Subjects at the Ordinary Level of the GCE, June 1971* (Occasional Publication 34). Joint Matriculation Board, Manchester, 1972.

III. Pairs of subjects

One of the most important assumptions of the method employing a test to investigate comparability of standards between subjects is that there is a demonstrable relationship between test performance and examination performance, and that these relationships are of the same order in all the subjects under consideration (see Appendix A, p. 93). In the research reported in Chapter II, these assumptions appear to have been validated, but if other methods could be devised that dispense with a test altogether, such methods would have greater appeal to those who believe that the aptitude test lacks face validity, and, perhaps more importantly, would obviate the need to spend thousands of pupil-hours in taking aptitude tests.

One such method that dispenses with an aptitude test was described in Chapter I: it is identified in this report as the subject-pair method and it was illustrated with the case of 7019 candidates from one CSE board who took both English and mathematics, obtaining an average grade in English of 2·96 and in mathematics of 3·55.

The subject-pair method

This chapter discusses the subject-pair method in relation to comparability of standards between the same ten subjects studied in Chapter II in GCE board 2. To illustrate the method in more detail the example of chemistry in GCE board 2 is again considered. There were 915 chemistry candidates in the sample and, from within this group, those candidates who were also taking art were identified. They numbered 134; their mean grade in chemistry was 6·19 and their mean grade in art was 5·11. This process was repeated for those candidates taking both chemistry and biology, both chemistry and English language, and so on. The results are shown in Table III.1.

In every comparison between the average grades in the two columns of Table III.1, the chemistry average grade is worse than the average grade in the other subject except in the case of French. The table also shows the means of both columns.

Figure 2 provides a diagrammatic representation of Table III.1, and also provides a similar diagrammatic representation for English language compared with each of the other nine subjects based on the results shown in Table III.2. The subject-pair points (squares) represent the mean of the means for chemistry and English language respectively.

Table III.1 The subject-pair method applied to the sample of GCE board 2 chemistry candidates

Subject	*Number of candidates*	*Mean grade in chemistry*	*Mean grade in the subject*
Art	134	6·19	5·11
Biology	486	5·41	4·39
English language	781	5·51	4·61
English literature	768	5·42	4·97
French	634	5·29	5·60
Geography	581	5·46	4·60
History	399	5·66	5·15
Mathematics	694	5·43	4·40
Physics	699	5·12	4·95
Mean (all subjects)	—	5·50	4·86

The estimate of severity or leniency in each subject is simply the difference between the two means of the means. Table III.1 shows that the mean of the chemistry mean grades is 5·50, while the mean of the other subject mean grades is 4·86. Chemistry in this sample thus appears as 0·64 grades severe (5·50 — 4·86 = 0·64). It should be noted that, in this method, the standards in each subject are being compared with the consensus standard of the other nine subjects. (In the methods using a test, described in Chapter II, the standards in each subject

Table III.2 The subject-pair method applied to the sample of GCE board 2 English language candidates

Subject	*Number of candidates*	*Mean grade in English language*	*Mean grade in the subject*
Art	659	5·14	5·30
Biology	1235	4·68	5·32
Chemistry	781	4·61	5·51
English literature	2245	4·95	5·41
French	1438	4·43	5·53
Geography	1659	4·90	5·56
History	1225	4·58	5·63
Mathematics	1650	4·81	5·40
Physics	833	4·60	5·39
Mean (all subjects)	—	4·74	5·45

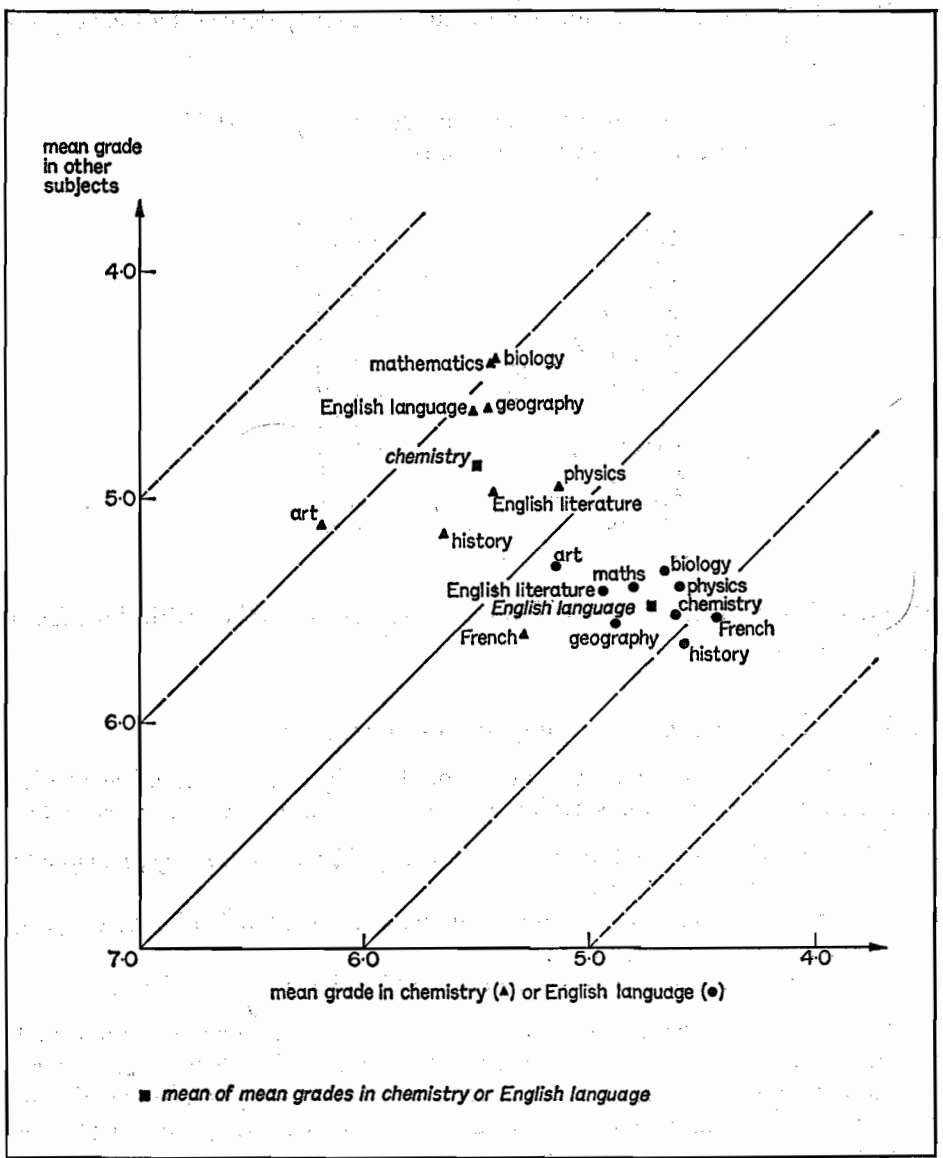

Fig. 2 *Subject-pair comparisons for GCE board 2*

are compared with the consensus standards of all ten subjects, i.e. the subject under study contributes to the consensus.) The whole problem of what is meant by a consensus is discussed in Chapter X.

Table III.3 presents the sample estimates of severity derived by the subject-pair method.

Table III.3 Sample estimates of mean grade severity in GCE board 2 (subject-pair method)

Subject	*Estimate of severity*
Art	−0·66
Biology	−0·17
Chemistry	0·63
English language	−0·70
English literature	−0·27
French	0·54
Geography	−0·08
History	0·21
Mathematics	0·01
Physics	0·50

Chapter VI compares the results from the subject-pair method with those from other methods.

THE ASSUMPTIONS OF THE SUBJECT-PAIR METHOD

In Chapter I, some of the assumptions of the subject-pair method were discussed briefly. Stated more formally, the major assumptions would appear to be:

a that candidates entered for each pair of subjects will, on average, be equally motivated in each subject;
b that the teaching of the candidates entered for each pair of subjects will, on average, be equally good;
c that the distribution of grades in each subject has the same shape.

Since the ten subjects under consideration are all popular and important subjects, the first assumption would appear to be reasonable. It might be argued that the observed leniency of English language (0·70 grades) was in part due to the candidates' greater motivation to do well in this subject, since a pass in English language is the most common entry requirement for employment and higher education. Mathematics is, however, nearly as common a required

qualification, but does not appear similarly lenient. It would therefore seem that differences between the motivation of candidates in different subjects are unlikely to be sufficient to account for the observed differences in attainment. Differences in motivation between less popular subjects and the major subjects might be of some significance.†

The validity of the second assumption is unlikely to be questioned. It should be noted, however, that teaching in this context refers not just to the quality of the teachers themselves but also to the quality of the teaching aids and equipment (e.g. are school geography rooms better equipped, on average, than school science laboratories?).

The third assumption also appears reasonable in the light of the evidence presented in Table III.4. The standard deviations of the grades in different

Table III.4 Sample means and standard deviations of grades in GCE board 2

Subject	*Mean grade*	*Standard deviation*
Art	5·27	2·31
Biology	5·35	2·34
Chemistry	5·45	2·39
English language	5·14	2·25
English literature	5·30	2·26
French	5·56	2·37
Geography	5·57	2·34
History	5·61	2·41
Mathematics	5·45	2·35
Physics	5·49	2·39

subjects vary very little (far less than do the mean grades) and results for other boards (not shown in this report) show a marked degree of constancy. Indeed, the very nature of the unofficial grading system used in GCE O level ensures that the distributions of grades in large-entry subjects have almost identical standard deviations and shapes.

These three assumptions have been discussed in the context of the subject-pair method since they are readily identifiable. It is the case, however, that all the different methods discussed in this report make these same three basic assumptions, but attention is not drawn to them again until Chapter X which provides a more detailed discussion of the necessary assumptions in all investigations in the field of subject comparability.

† See also *A Comparison of Awards by Subject Panels for the period 1966–1970:* Research Report 10 of the Yorkshire Regional Examinations Board (YREB Harrogate, 1970).

IV. All subjects taken together

The previous method, the subject-pair method, was seen to be a method of investigating the comparability of standards between subjects which did not require the use of an external reference test. The purpose of this chapter is to explain another such method, developed independently, which takes more than one pair of subjects at a time into account.

The average GCE O-level candidate at 16+ offers about five subjects. The grade obtained by a particular candidate in a particular subject tells us little about the candidate's overall performance and nothing about the subject's severity. Given more grades for a candidate, we begin to build up an overall picture of the capability of that candidate; given more candidates, with all their grades, the picture of the relative comparability of standards in the subjects emerges.

One of the fundamental assumptions of the subject-pair method is that the distribution of grades in all subjects has the same shape. This assumption is also vital to the method presented below.

The basis of the proposed method is that of using a reference to gauge the ability of groups of candidates attempting different subjects. The reference used is denoted by UBMT, which is defined below (the origin of these letters relates to a consideration of the unbiased mean total, but it will be appreciated that the UBMT reference as defined is not a total in the accepted sense of the word). If chemistry is the subject under consideration, the UBMT for any candidate taking chemistry is simply the sum of the grades he obtains in all the other subjects he attempts, divided by the number of those subjects: more simply it is the 'mean grade in all *other* subjects attempted'. For all candidates taking a subject, e.g. chemistry, UBMT is the mean grade of all other subjects attempted for all candidates taking chemistry. As such it is not the mean of the individual candidate's UBMTs, but the weighted mean (the weights being the number of other subjects attempted). It will be appreciated that candidates attempting single subjects are not eligible for inclusion in this type of analysis as they contribute nothing to UBMT (no 'other' subjects).

In Table IV.1, the mean grade for those candidates taking chemistry with at least one other subject (914 candidates in all) is 5·44. However, these 914 candidates obtained an overall mean grade of 4·89 in the other subjects they attempted.

Taking the reverse nature of the grade scale into account, it is possible to say that on average chemistry was 0·55 grades severe (5·44 — 4·89) since it would

Table IV.1 Sample estimates of mean grade severity and their derivations in GCE board 2 (UBMT method)

Subject	*No. of candidates*[a]	*Mean grade in subject* (MG)	*Mean grade in other subjects attempted* (UBMT)	*Severity*[b] (MG — UBMT)
Art	734	5·27	5·90	−0·63
Biology	1351	5·32	5·25	0·08
Chemistry	914	5·44	4·89	0·55
English language	2741	5·09	5·72	−0·64
English literature	2533	5·29	5·51	−0·22
French	1633	5·56	5·04	0·52
Geography	1916	5·55	5·46	0·09
History	1413	5·60	5·29	0·31
Mathematics	1884	5·44	5·32	0·12
Physics	972	5·42	4·93	0·50

[a] Total number of candidates attempting 2 or more subjects = 3114.
[b] Rounding error accounts for some discrepancies.

have been expected that these two mean grades (MG and UBMT) would have been equal had comparability between subjects existed. The other subjects may be treated in a similar way and the results are presented in Table IV.1, the final column showing some subjects to be over a half a grade severe and others over a half grade lenient. Figure 3 shows these results graphically. The broken line has been drawn such that at all points on it the mean grade is equal to UBMT; subjects falling on – or very near – this line may therefore be regarded as neither severe nor lenient.

In essence the UBMT approach to subject comparability is easy to adopt and it is easy to see what is happening; there are, however, some points to be noted. The most important is that of a consensus. For each subject considered in turn, the associated UBMT value does not, by definition, contain a contribution from that subject and this leads to a fair criticism of the method as a whole. Put another way, the UBMT scale is not the same for each subject. It will also be appreciated that different candidates offered not only different numbers of subjects, but a different 'mix' of subjects: some had a bias towards science subjects, and some towards arts subjects. Table IV.2 gives the overall distribution of the number of candidates attempting a given number of subjects.

A consequence of using the UBMT method is now apparent: depending on the subject under consideration, not only will that subject be excluded from the calculation of UBMT, but the subjects common to the calculation of UBMT

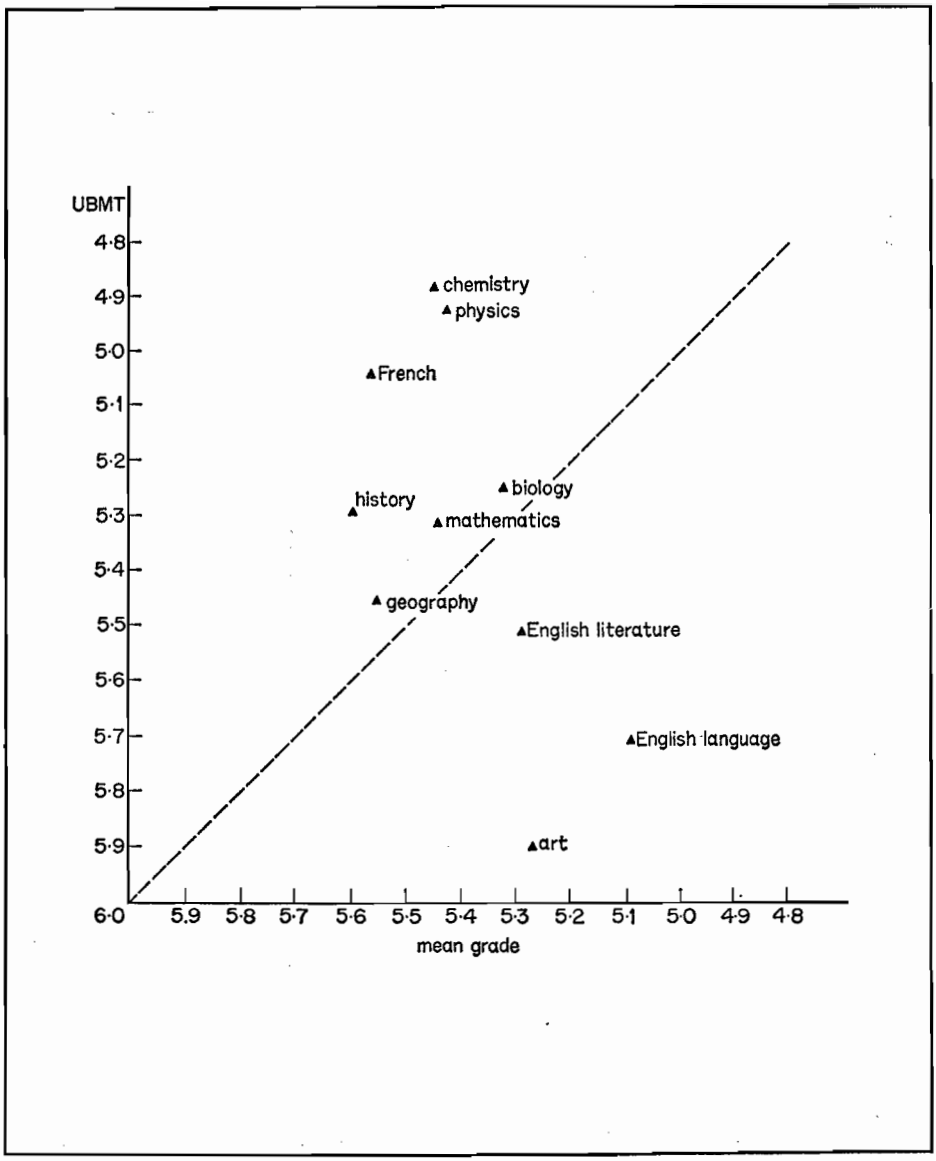

Fig. 3 *Graph of* UBMT *against mean grade for GCE board 2*

Table IV.2 Numbers and percentages of candidates attempting a given number of subjects in GCE board 2

No. of subjects attempted[a]	*No. of candidates*	*% of candidates*
1	(211)[b]	—
2	297	9·5
3	395	12·7
4	425	13·6
5	582	18·7
6	592	19·0
7	496	15·9
8	266	8·5
9	60	1·9
10+	1	—
Total	3114	100

[a] Mean number of subjects attempted = 5·17.

[b] As mentioned above, those candidates offering a single subject could not be included in the analysis. Thus, out of a total number of candidates of 3325, 211 only offered a single subject; they were therefore dropped from the analysis to leave a sample of 3114 candidates.

for, say, geography and physics may well be included in different proportions. If all the subjects were strictly comparable this would not matter in the least; as this does not appear to be the case, it is important to be aware of the shortcomings of the method. While the data on the relative popularity of the other subjects attempted could be obtained it is difficult to see to what use it would be put. Instead, Table IV.3 simply presents, by subject, the mean number of subjects attempted.

It may be seen that those candidates taking chemistry, physics and French are attempting, on average, over one subject more than those attempting English language and literature. In general, the subjects attempted by candidates attempting a large number of subjects overall tend to be severe. It is not likely that there is any degree of causality to be implied here since it is probably the nature of the subject concerned which affects its severity and attracts candidates attempting large numbers of subjects.

An important point to be discussed concerns the estimates of severity obtained. In most methods, the mean severity of a group of subjects is zero but this is not the case with the UBMT method; a glance at the severity column of Table IV.3 will show immediately that there is a mean tendency towards severity over all

Table IV.3 Mean number of other subjects attempted and estimates of severity in GCE board 2

Subject	*No. of candidates*	*Mean no. of subjects attempted*	*Severity*
Art	734	5·54	−0·63
Biology	1351	5·99	0·08
Chemistry	914	6·66	0·55
English language	2741	5·28	−0·64
English literature	2533	5·44	−0·22
French	1633	6·26	0·52
Geography	1916	5·77	0·09
History	1413	5·94	0·31
Mathematics	1884	5·96	0·12
Physics	972	6·48	0·50
Overall	3114	5·17	—

ten subjects. The reason for this tendency lies in the above argument regarding the total number of subjects attempted and the relative popularity of the subjects. By allowing all the other subjects attempted by a candidate to be included in the calculation of UBMT for a particular subject, the more popular subjects will appear in these calculations more often than the unpopular ones; in Table IV.3, the two most popular subjects (English language and English literature) are both on the lenient side. These two subjects will appear in the UBMT calculations for most other subjects and this will, therefore, result in the calculated values of UBMT for most subjects being on the low side. The mean grade minus UBMT will, as a result, be on the high (severe) side.

Thus, because of the way the 'reference ability' of candidates attempting individual subjects is calculated, the effect of using the UBMT method is to arrive at estimates of severity which in this case are 'too severe' overall. This raises the problem of what is meant by a consensus standard (see Chapters III and X). However the weighted sum of the severity estimates, using the numbers of candidates attempting a subject as the weights, is zero, as this allows for the popularity of a subject to be accounted for.

The definition of a consensus standard by the UBMT method is therefore seen to be such that the mean of the weighted estimates of severity is zero, rather than the unweighted mean. Put another way, the overall standard is defined by the average candidate performance, rather than the average subject performance. A fuller discussion of this topic will be presented in Chapter X. In the meantime, realizing that a choice of origin of standards (consensus) is largely arbitrary (in

the case presented one could argue that all subjects are severe relative to English language and art or that all subjects are lenient relative to chemistry) it is quite possible to alter the values of severity obtained by the UBMT method such that the unweighted sum is zero.

In Table IV.4, it may be seen that the sum of the severities is 0·68 grades.

Table IV.4 Sample estimates of mean grade severity in GCE board 2 (UBMT method)

Subject	*No. of candidates*	*Original severity estimates*	*Severity estimates adjusted to mean zero*
Art	734	−0·63	−0·70
Biology	1351	0·08	0·01
Chemistry	914	0·55	0·48
English language	2741	−0·64	−0·70
English literature	2533	−0·22	−0·29
French	1633	0·52	0·45
Geography	1916	0·09	0·02
History	1413	0·31	0·25
Mathematics	1884	0·12	0·05
Physics	972	0·50	0·43
Unweighted sum	—	0·68	0·00

Subtracting this amount in equal proportions across all ten subjects (0·068 grades off each value in the second column) gives the results in the final column of the table. The sum of these adjusted UBMT values is now zero.

A further discussion of these topics in relation to the other methods proposed for estimating the degree of comparability of a subject will be found in Chapter VI.

V. Analysis of variance

One of the major difficulties of subject comparability is that one has to deal at one and the same time with possible variations of severity in the grading of the different subjects, and with the sure knowledge that candidates differ very considerably in ability. A further complication is that the calibre of candidates may vary from subject to subject. It was to deal with this difficulty that resort was made to a well-known statistical technique called the analysis of variance. The actual computations involved in this work are complicated, and the reader is spared an explanation of their details (see below for references), but the underlying idea is a simple one.

The method that has been used makes the assumption that the grade obtained by a given candidate (candidate i) in a given subject (subject j) may be expressed as the sum of four parts:

- a which may be regarded as an average grade for all candidates and subjects, although it may differ somewhat from the mean;
- c_i which depends only on the candidate concerned, and which may be thought of as the overall ability of candidate i to obtain good grades in the examination;
- q_j which depends only on the subject concerned, and which may be described as the severity of the grading of subject j;
- e_{ij} which depends both on the candidate and the subject, and which may be thought of as the error in estimating the grade of candidate i in subject j from the three terms above.

Thus

$$(\text{grade of candidate } i \text{ on subject } j) = a + c_i + q_j + e_{ij}$$

One great advantage of this method is that the calibre of the candidates offering the different subjects and the severity of the grading of the subjects are dealt with simultaneously by the analysis, thus overcoming the difficulty mentioned in the first paragraph. Bearing this in mind, it may be helpful to explain the relation of the statistic c_i to the candidate's grades. Supposing the severities q_j of the different subjects are known, the appropriate one is subtracted from each grade obtained by the candidate, and the mean value calculated. For example, if a candidate obtains grades 3, 4, 5, 3, in subjects with severities 0·1,

0·2, −0·1, −0·3, the value for this candidate of $a + c_i$ will be

$$\frac{(3 - 0{\cdot}1) + (4 - 0{\cdot}2) + (5 - (-0{\cdot}1)) + (3 - (-0{\cdot}3))}{4} = 3{\cdot}775$$

which differs by only 0·025 from the candidate's mean grade. This means that $a + c_i$ may be thought of as the candidate's mean grade corrected for the severities of the subjects offered. It is found that c_i correlates well with grades in the various subjects: in board 2 the mean correlation over ten subjects was 0·710 while the scholastic aptitude test had a mean correlation of 0·357.† This suggests that c_i may be regarded as a satisfactory measure of the overall ability of a candidate to perform well in the examination.

However, the chief interest in this report is in the values of q_j. This statistic is the same for all candidates offering subject j; if q_j is larger than average, grades of all candidates in this subject will tend to be larger than average, which means that their results may be interpreted as being worse than average. This suggests that q_j may be regarded as the severity of the grading of subject j. The phrase 'will tend to be larger than average' was used because it is possible that a group of high ability might attempt a subject – for instance Latin – and their ability might compensate for the severity of the grading and so result in grades near the average.

Advantages and disadvantages

In common with other methods using only the grades awarded by the board concerned, the method is cheap in that all the extra work for the analysis is to be found in taking a representative sample and punching cards for a computer. It might well be employed as a standard research tool within a board to monitor severities of subjects year by year and to establish a basis for moderating options in a given subject when there are no questions common to the options concerned. Another advantage of the method is that further statistical techniques exist for testing the significance of differences between pairs of severities.

On the other hand, the fact that the method uses only grades from one board at a time means that severities are not strictly comparable between boards. This is largely because there is no direct way of deciding whether the points of zero severity of different analyses are comparable. If the method is applied to the two sexes in a given board, this is still strictly the case but one would have more confidence that a comparison was meaningful. However, it is still possible to compare the rank order of the severities of two boards, bearing in mind that

† This is an overall correlation, not quite identical with the pooled within-school correlations shown in Appendix B, p. 99.

small differences within a board may be reversed if another sample is taken from those who took the examination in the year concerned. If the samples are representative, it is thought likely that the consensus of the samples would be close. Another comparison which would be of interest is that of the severities of the same board in different years. The next section presents another aspect of the same problem.

Number of subjects in the analysis

In common with other methods, a decision has to be made on how many subjects to include in the analysis. It is therefore of interest to know whether the analyses carried out with different numbers of subjects will give the same values for the subject severities. Clearly it is reasonable to include those subjects most often offered and, as the subjects were roughly arranged in this order on the punched cards, the decision amounted to how many to include. Ten subjects had been taken by Forrest[1] so this has been taken as the general rule in this report, but two subjects were added to see whether the severities were affected. The results for board 2 are shown in Table V.1 and give little cause for disquiet: the ten subjects included in both analyses come out in the same order, the average difference between the corresponding severities being about 0·07. This difference arises because the sum of the q_j for each analysis is zero, and the sum of the severities for RE and Latin is 0·67, so that −0·67 has to be distributed among the other subjects in the second analysis. This provides a good illustration of

Table V.1 Sample estimates of mean grade severity in GCE board 2 (ANOVA method)

Subject	*10-subject analysis*	*12-subject analysis*
Art	−0·69	−0·76
Biology	−0·04	−0·08
Chemistry	0·53	0·44
English language	−0·64	−0·70
English literature	−0·29	−0·35
French	0·43	0·36
Geography	−0·01	−0·07
History	0·21	0·15
Latin	—	1·26
Mathematics	0·04	−0·03
Physics	0·46	0·37
Religious education	—	−0·59

two analyses with different points of zero severity. A similar trial was carried out with board 6 and the order of the severities remained the same except for mathematics and chemistry, their severities changing from 0·14 and 0·15 to 0·19 and 0·18.

Another possible variation in the analysis is to exclude candidates who have only offered one or two subjects. Clearly, if a candidate has offered only one subject, his one grade can throw little light on to the question of subject comparability. And, in so far as the whole business is concerned with establishing a consensus across subjects, it seems reasonable to suppose that this can be done most effectively by examining the grades of candidates who have offered a fair number of subjects. Accordingly, another analysis was carried out on board 6, eliminating candidates who offered less than four subjects. RE and Latin were included and there were two changes in rank order: geography and history changed from 0·08 and 0·05 to 0·13 and 0·15, while biology and mathematics changed from 0·17 and 0·14 to 0·22 and 0·18. These, and the results above, although by no means conclusive, suggest that the rank order of the difficulties obtained by the analysis of variance method is fairly stable. In general the analysis has been run on the ten standard subjects.

Differences between severities

It is clear that, had different samples been obtained from the boards, the values of the severities would not have been precisely the same, so it is natural to ask what reliance may be placed in the values shown. Unfortunately the samples were not random ones and so the results obtained from applying the tests of significance described below must be treated with reserve. However, in any further investigation it would be highly desirable to apply such tests and accordingly they are described briefly and illustrated by the severities of board 2.

The first hypothesis tested is that severities in the population are equal. Since the sum of the severities in any analysis is zero, this is equivalent to the hypothesis that all the severities in the population are zero. This means that the values obtained would have arisen through chance – and that they might equally well change their signs if another sample were taken. This hypothesis is tested by what is known as an F-test; the value obtained for F was 67·2, while the critical value for significance at the 0·1 per cent level is about 3. This means that the hypothesis is rejected with very great confidence and it is concluded that the severities in the population of board 2 are not all zero.

Having established beyond reasonable doubt that there are non-zero severities in board 2, it is natural to inquire whether individual severities in the board are zero or not. However, this is not a question that can be answered because the severities can only be interpreted *in relation to each other*. This has already been

seen in Table V.1 (which illustrated how the values of the severities were affected by the introduction of two extra subjects in the analysis), but it is also implicit in the analysis itself since the sum of the severities in any analysis is zero – which is equivalent to saying that there is no predetermined origin from which we can measure severities. On the other hand, it is possible to test for differences between the forty-five pairs of severities by means of what is known as an *S*-test. When this has been done, there remains the problem of studying the results and for this purpose it is convenient to use what we term a grid diagram (see Figure 4). The severity of each subject is plotted along each axis; for clarity, the axes are not shown but the scale is indicated below the diagram. By this means we have a ready way of comparing the forty-five pairs of severities between the different subjects. The grid diagram also gives a general picture of the levels of significance of the differences between pairs of subjects. Significance at the 5 per cent level is indicated by a triangle, the 1 per cent level by an open circle, and the 0·1 per cent level by a dot. These signs are placed at the two points of the grid corresponding to that pair of subjects.

As stated above, the results must be treated with considerable reserve but in so far as most of the significance levels are high, it is thought that the overall picture is not misleading, the more so as results from other boards (not shown) are similar.

Assumptions and technicalities

The method of analysis used to obtain the results in this chapter and Appendix C is described by Backhouse in *British Examinations: Techniques of Analysis*,[2] where the normal equations will be found. Tests of significance used on the results are described in section 4.4 of Scheffé's *The Analysis of Variance*;[3] they consist of a test of the hypothesis that all the severities are zero under the hypothesis of no interactions, and the *S*-method applied to each pair of severities within the analysis. The assumptions involved are that the errors are uncorrelated and have equal variance, and that the grades have a joint normal distribution. It has not been investigated how far these assumptions are violated, but high values of F were obtained when testing the hypothesis that all the severities are zero. The values for the four boards which have received most attention were 65·8, 67·2, 21·4, 38·5, so it may be stated with considerable confidence that there are real differences in the severities of the subjects in respect of the samples investigated. About the confidence intervals obtained by Scheffé's method there is less certainty, and they have not been quoted.

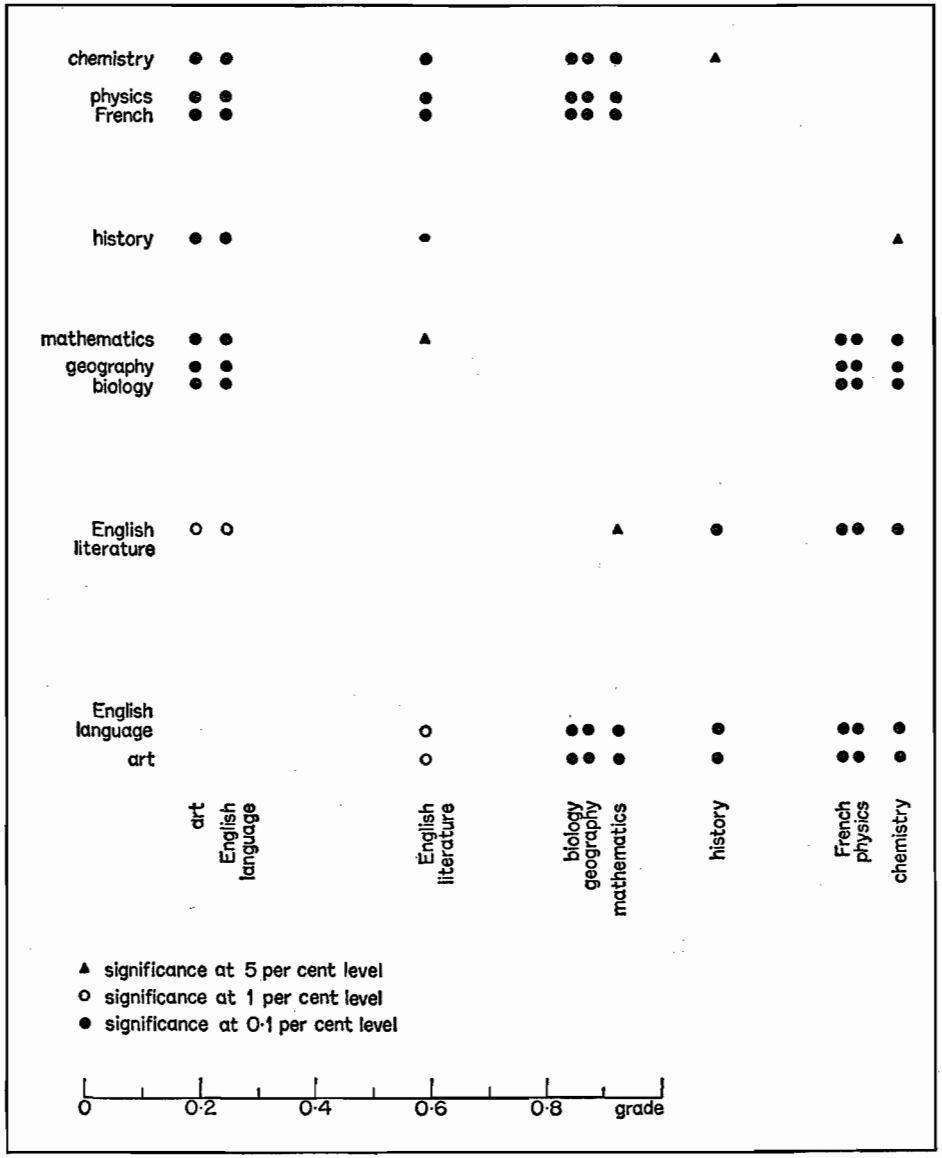

Fig. 4 *Grid diagram indicating significance levels of differences in severities between pairs of subjects in GCE board 2*

References

1. G. M. FORREST, *Standards in Subjects at the Ordinary Level of the GCE, June 1970* (Occasional Publication 33). Joint Matriculation Board, Manchester, 1971.
2. J. K. BACKHOUSE, 'Reliability of GCE examinations: a theoretical and empirical approach' in *British Examinations: Techniques of Analysis* by D. L. NUTTALL and A. S. WILLMOTT. NFER, Slough, 1972.
3. H. A. SCHEFFÉ, *The Analysis of Variance.* Wiley, New York, 1959.

VI. The methods compared

The reader has now been presented with no less than five methods of tackling the problem of comparability between subjects, two of which depend on the use of an aptitude test, and three of which do not. Of course, it may not be accepted that there is any true lack of comparability in the present situation and that the 'differences' found between subjects are inherent differences. Leaving that point aside until later (Chapters VIII and X), it would now seem reasonable to compare the proposed methods. This comparison will cover the theoretical differences and similarities, their potential for use, and the results obtained by using the different methods. It should be noted that in all the comparisons in this chapter, the UBMT values of severity have been adjusted to give an unweighted mean of zero (see Chapter IV, p. 37) in order that meaningful comparisons of results across methods may be made.

The first step in comparing the five methods is to collate the various estimates of severity for the subjects in GCE board 2 that have been presented in the previous chapters, and these results are presented in Table VI.1.

It may be seen that there is a general tendency for the five methods to agree on the order of the severities for the subjects, although some discrepancies are apparent (for example, by regression, history is more severe than mathematics,

Table VI.1 Sample estimates of mean grade severity for GCE board 2: the five methods compared

	METHOD				
Subject	*Regression*	*Guideline*	*Subject-pair*	UBMT	ANOVA
Art	−0·49	−0·80	−0·66	−0·70	−0·69
Biology	−0·14	−0·19	−0·17	0·01	−0·04
Chemistry	0·33	0·62	0·63	0·48	0·53
English language	−0·49	−0·57	−0·70	−0·70	−0·64
English literature	−0·24	−0·32	−0·27	−0·29	−0·29
French	0·25	0·29	0·54	0·45	0·43
Geography	0·09	−0·05	−0·08	0·02	−0·01
History	0·16	0·04	0·21	0·25	0·21
Mathematics	0·12	0·18	0·01	0·05	0·04
Physics	0·37	0·72	0·50	0·43	0·46

whereas the situation is reversed when the guideline estimates are used); the actual magnitude of the values also varies considerably.

There is a more general agreement within the last three methods (all based on the same 'internal' evidence) than there is between these methods and those based on the use of an aptitude test. There are also considerable discrepancies within these two 'test' methods and because of this, and the fact that the administration of a test of any kind requires a special exercise to be carried out in addition to the routine examining procedure, the results of the regression and guideline methods will only be treated lightly in this chapter. The main emphasis will be placed on the use of methods that may be adopted without the use of any test outside the normal examination.

Before leaving the use of a reference test, there is one major point which should be considered. The three internal methods of estimating severity all deal with a different consensus standard but, whatever consensus is adopted, it is impossible to say that the consensus in one board is of the same standard as the consensus in another board without external evidence of some kind. Whether or not the reference test is used *within* a board to investigate comparability between subjects, it (or something like it) is certainly needed to relate the standards in one board to those in another board. This point and a discussion of the different nature of the consensus standard implied by the three internal methods will be discussed further in Chapter X.

Tables C.1, C.2 and C.3 (pp. 106–7) give a further comparison of the five methods by presenting severity estimates for the ten subjects within GCE boards 1, 3 and 4. In each case the general pattern of findings from board 2 are replicated. Before making too many actual comparisons, it is constructive to look at the number of candidates in each of the four boards. Table VI.2 gives the numbers of candidates taking each subject by board as used for the regression, guideline and ANOVA† methods; the bracketed figure in each case represents the numbers of candidates as used for the subject-pair and UBMT methods. As has been seen, neither of these methods can make use of data from candidates attempting a single subject only, so that the discrepancy in each case is the number of candidates who only attempted that particular subject. As one might expect, these differences are larger in the less 'selective' subjects.

In order to make comparisons between the subject-pair, UBMT and ANOVA methods more easily, the magnitude of the difference between each pair of estimates has been calculated for each subject in each board. Table VI.3 gives the average differences (across all ten subjects) found in the four GCE boards.

† Analyses reported in Chapters V and VII show that dropping candidates attempting single subjects does not materially alter the severity estimates for the ANOVA method in the two cases studied.

Table VI.2 Numbers of candidates in analysis, by GCE board

Subject	GCE BOARD 1		2		3		4	
Art	562	(464)	765	(734)	174	(159)	269	(249)
Biology	837	(818)	1372	(1351)	289	(287)	560	(559)
Chemistry	449	(448)	915	(914)	153	(153)	334	(331)
English language	1579	(1519)	2828	(2741)	591	(546)	798	(742)
English literature	1308	(1303)	2548	(2533)	502	(501)	612	(612)
French	928	(925)	1635	(1633)	342	(342)	608	(606)
Geography	972	(966)	1948	(1916)	448	(430)	634	(630)
History	740	(735)	1422	(1413)	270	(267)	666	(662)
Mathematics	1091	(1065)	1892	(1884)	476	(459)	689	(671)
Physics	562	(544)	977	(972)	179	(179)	399	(398)
Total no. of candidates	1912	(1671)	3325	(3114)	756	(655)	1223	(1114)

It will be noted that the differences between the methods in boards 3 and 4 are, on the whole, larger than those found in boards 1 and 2, particularly in comparisons involving the subject-pair method. The reason for this almost certainly lies in the sample sizes in each of the four boards (see Table VI.2). Boards 1 and 2 both have larger sample sizes than boards 3 and 4. As there may be, in board 3, say, only a few candidates in the sample taking both chemistry and art, as opposed to much higher numbers of candidates taking, say, mathematics and English language, the mean grades of these candidates will not necessarily be wholly representative of all those candidates in board 3 taking

Table VI.3 Mean differences, across 10 subjects, between the estimates of severity given by the subject-pair, UBMT and ANOVA methods (GCE grades)

Board	METHOD *Subject-pair*/UBMT	*Subject-pair*/ANOVA	UBMT/ANOVA
1	0·08	0·05	0·04
2	0·07	0·05	0·03
3	0·09	0·09	0·03
4	0·10	0·10	0·05
Over all 4 boards	0·08	0·07	0·04

these two subjects (i.e. in the population). The errors in the estimation of the severity of both chemistry and art are therefore likely to be adversely affected. It should be noted that Chapter VIII tackles the problems of sample sizes and the effects of using the various methods on populations and samples of candidates. Over all boards, then, the UBMT and ANOVA methods agree very well, followed by the subject-pair and ANOVA methods and the subject-pair and UBMT methods.

A point now to be considered concerns the nature of the proposed methods of estimating severity. Two aspects of each method are of considerable importance to discuss: the question of a weighted or unweighted approach and the question of 'bias'.

These two concepts – bias and weighting – are important to understand, at least in theory. It is also important to appreciate that there are two areas in which each may operate: that of calculating severity estimates, and that of defining a consensus standard (which will be held over for discussion in Chapter X). It will have been appreciated that, in order to compare the relative performance of different groups of candidates attempting different subjects, some measure of candidate performance was required. The question of bias hinges on whether or not the subject under consideration should contribute to the calculation of that measure of performance; if it does, the method is biased, if not, the method is unbiased. The question of weighting brings in quite another concept: should a subject with twice as many candidates entering for it as another subject contribute twice as much to the calculation of the consensus as the second subject? If this is the case, then the method is weighted; if each subject contributes equally then the method is unweighted.

A clarification of this situation may be seen as follows. Suppose that an analysis were carried out on a group of candidates by both a weighted and an unweighted method. Suppose, also, that a number of candidates were then added who were representative of certain aspects of the main group: they may have entered only a few of the subjects represented in the main group, but their performance in each subject was representative of that found in the main group. If an analysis were then carried out on the combined set of candidates, a weighted method of estimation of severity estimates would give rather different results, whereas an unweighted method would give the same results. The difference lies in whether it is the mean candidate performance or the mean subject performance that is being considered as a criterion. It should be borne in mind that although the subject-pair method is unweighted and the UBMT method weighted, the subject-pair method could easily be a weighted method. Table III.1 (p. 28) shows, for chemistry, the number of candidates involved in each subject-pair. The means at the foot of the two right-hand columns (5·50 and 4·86) have been arrived at by finding the unweighted average of the subject mean grades in these

columns. Thus, the mean chemistry grade (5·50) is *not* the mean grade of all those candidates (taking more than one subject) taking chemistry (cf. the UBMT method). If weights (the number of candidates in a subject-pair) were used, then the mean at the foot of the right-hand column would become the UBMT value for chemistry in board 2. The new mean at the foot of the middle column would not, however, be the mean grade in the subject as usually defined. This is seen by noting that a candidate taking, say, five subjects in addition to chemistry would have his chemistry grade included five times in the weighted total. This mean would then tend to be rather lower (imply a higher standard) than the usually defined sample mean grade – on the basis that the more subjects candidates attempt, the better their mean grade. Thus it may be seen that the subject-pair method and the UBMT method have elements in common but are not simply weighted and unweighted versions of the same method. In terms of bias the

Table VI.4 The relationship between the methods of estimating severity in terms of biased/unbiased and weighted/unweighted attributes

	Biased	*Unbiased*
Weighted	ANOVA	UBMT
Unweighted	Regression (Guideline)	Subject-pair

issue is clear: both methods are unbiased as the grade of candidates in the subject under consideration is not included in the 'reference ability'.

It is now possible to consider the ANOVA model; where does it fit into the structure presented above? In Chapter V the description of the ANOVA model states that the term c_i represents the measure of a candidate's ability to obtain good grades in an examination and that the term q_j represents the subject severity. The data are used in the estimation of the c_i and the q_j in a manner such that the estimates of severity turn out to be both biased and weighted. Candidates taking one subject need not be removed from the analysis.

Finally the regression method (and also the guideline method) use information within subjects to form biased but unweighted estimates of severity (see Appendix A, p. 93). Table VI.4 gives the relationship between the various methods.

A glance at the table above is enough to indicate that it is very unlikely that the methods will ever yield identical severity estimates as they are based on rather different basic philosophies.

From the point of view of the present *results*, all five methods give very similar

values of severity and the biased/unbiased, weighted/unweighted aspects are seen to be very largely unimportant in affecting these results. Consideration should be given, however, to locating those situations where different results could be obtained if one method was used in preference to any other. It would appear that the question of whether or not a weighted method should be used is of considerably greater importance than the question of whether or not a biased method should be used; that this is so stems from the unreliable nature of statistics based on small sample sizes. The effect of such small sample sizes will be most acute in methods such as the subject-pair method and it is obvious that this method should be avoided in such situations. If the sample of candidates in the analysis for some reason only attempt a relatively small number of subjects on average, then the UBMT method (as well as the subject-pair method) is likely to become rather unstable, as there will be a small number of 'other' subjects attempted.

It is undoubtedly the ANOVA method which is the most versatile and the most likely to yield sensible results with small samples and low numbers of subjects attempted. (The variability of the results is still dependent on the sample size, however, and all that may be said is that this variability is likely to be smaller with the ANOVA method than with the other two internal methods.) The ANOVA method, however, needs relatively sophisticated computing techniques in order to work at all. The subject-pair and UBMT methods are undoubtedly simpler in operation and concept and could be done by hand, especially the latter. It is at this stage that we should move on to a discussion of the sampling properties of these three methods, leaving the philosophical issues until Chapter X.

VII. The methods applied to one CSE board

While the basic material for this report was being collated, contact was maintained with the examining boards in both the GCE and CSE sectors as part of the work of the NFER projects concerned. Through this contact it became clear that many of the boards were becoming increasingly concerned with what we have called here 'subject comparability' and that there was a general desire to learn more about what is happening in examinations at 16+. Part of this feeling was expressed by one CSE board in making available to the Foundation a complete breakdown of their summer 1971 entry, by subject entered and grade awarded. For over 21 000 candidates, the data offered consisted of all Mode I subjects entered by each candidate together with the results obtained in each subject (CSE grade). These data were of considerable interest and they were used with the distinct aims of investigating certain areas of subject comparability which could not otherwise have been considered.

Aims of the investigation

The first aim was to look at the sampling techniques currently adopted in much of the authors' current work. Where a sample of, say, 1000 candidates is required from an entry of 8000 candidates in all, a statistically correct procedure would be to pick a 12·5 per cent sample completely randomly, each candidate in the population having an equal chance (1 in 8) of being included in the sample. This process, while ideal in nature, is extremely time-consuming in practice. The procedure adopted in the past has been one of choosing, in the example given above, every eighth candidate; this implies that the candidates are ordered in some way – by numerical order, or by alphabetical order – and that once the first candidate is chosen, the sample is defined. In such a situation it is not always clear that a good sample is drawn, although there is no reason to suspect that there will be any systematic bias unless the candidates are in a rather peculiar order initially. In the case of the CSE board, a 10 per cent sample was drawn on an 'every tenth candidate' basis and analysed independently from the population. The results of these analyses are given on pages 53–4.

The second aim of analysing the population data was to investigate the stability of the proposed methods of estimating a subject's severity from a sample of candidates, when compared with the values obtained using the same method for the population of candidates from which that sample was drawn. This aim

is obviously dependent upon the acceptance of the '1 in *N*' sampling procedure just mentioned.

The third aim of using the population data was to look at minority subjects. It will be appreciated that if a subject is taken by relatively few candidates, by the time a sample is taken those entries are very small indeed. In the present case, for example, only 351 candidates took German CSE with the CSE board in summer 1971 (see Table VII.2, page 54). In the 1 in 10 sample only thirty-nine candidates appear, which makes the estimation of the severity for German subject to considerable error. In the population, however, it is possible to use all the 351 candidates, and so obtain a much better estimate of the severity for German.

Having broadly stated the reasons for being so interested in the data for the population of CSE candidates in one board, it is only reasonable to comment on two other aspects of the results which would be available from the analysis of such a population. Initially there is the fact that when dealing with a sample of candidates, no matter how carefully drawn that sample is, the results are always open to the criticism that with a different sample of candidates different results would have been obtained. This difference could be reduced to a minimum by careful sampling but would always exist. (It should be noted that the severity estimates in two distinct random samples are very likely to be numerically different, but very unlikely to be other than statistically equivalent.) When dealing with a population of candidates, no such criticism can apply – the values of severity obtained are values of *real differences* as defined by the method of analysis adopted.

Secondly, the following chapter compares the results – as opposed to the methods – of subject comparability investigations in three ways; one of these ways is across sectors (GCE and CSE) and in the 1968 comparability exercise the number of candidates in each CSE board was only about 400. Since the analysis of such a relatively small number of candidates would yield severity estimates which were likely to be somewhat unstable, the opportunity of comparing the population results from one CSE board with the results in the GCE sector was not to be missed. Finally, of course, the data were very much more up to date (1971) than the bulk of the data with which this report is concerned (1968).

The results

In order to see whether the method of sampling used yielded a representative sample, it is possible to look at the numbers of candidates attempting various subjects and the mean grade obtained both for the population and also for the 1 in 10 sample. There is a different approach adopted here from that described

previously in that twenty subjects are now included in the analysis as opposed to ten. It should be appreciated that there were, in all, twenty-seven Mode I subjects offered by candidates in the CSE board but, in order to keep a manageable number of subjects, only the most popular twenty were chosen.

The first comparison is between the numbers and percentages of candidates offering a given number of subjects, and this information is presented in Table VII.1.

Table VII.1 Comparison of the number and percentage of candidates entering a given number of subjects in a 1 in 10 sample and the population for the CSE board

	No. of candidates		*% of candidates*	
No. of subjects[a]	*in sample*	*in population*	*in sample*	*in population*
0	13	180	0·6	0·9
1	402	4071	19·7	19·9
2	264	2681	12·9	13·1
3	270	2592	13·2	12·7
4	292	3159	14·3	15·4
5	377	3540	18·5	17·3
6	265	2586	13·0	12·7
7	126	1181	6·2	5·8
8	28	355	1·4	1·7
9	4	67	0·2	0·3
10	—	4	—	—
Total	2041	20 416	100	99·8

[a] Mean number of subjects attempted: 3·69 in the 1 in 10 sample, 3·66 in the population.

The first point to note is that the 13 candidates in the sample and 180 candidates in the population apparently offering no subjects at all reflect the fact that only twenty out of a total of twenty-seven subjects were considered in the analysis. The omitted subjects were civics, engineering science, home economics (home-making), music, needlework (embroidery), rural studies and science of living. Table VII.1 shows that less than 1 per cent of candidates (0·6 and 0·9 in the sample and the population respectively) attempt one or more of these seven subjects only. They are therefore ignored for the purposes of this analysis.

A quick look at Table VII.1 shows that the proportions of candidates entering a given number of subjects are generally very consistent between the sample and the population. A slight disturbance occurs for those candidates offering

four and five subjects, as the sample is under-represented by those offering four subjects (14·3 per cent and 15·4 per cent) whereas the sample is over-represented by those offering five subjects (18·5 per cent and 17·3 per cent). The mean number of subjects offered is very similar, although slightly higher for the sample, as may be expected from the disturbance just noted.

Secondly, Table VII.2 gives the number of candidates attempting each of the twenty subjects, together with their mean grade, for the population and sample.

Table VII.2 Comparison of the numbers and percentages of candidates entering each of twenty subjects and the mean grade obtained in a 1 in 10 sample and the population for the CSE board

	NO. OF CANDIDATES (% *of total*)					MEAN GRADE		
Subject	*Sample*		*Population*		*Difference*	*Sample*	*Population*	*Difference*
Arithmetic	239	(11·71)	2426	(11·88)	−0·17	4·00	4·06	−0·06
Art and crafts	404	(19·79)	3851	(18·86)	0·93	3·40	3·34	0·06
Biology	232	(11·37)	2393	(11·72)	−0·35	3·21	3·33	−0·12
Business studies	110	(5·39)	1027	(5·03)	0·36	3·39	3.42	−0·03
Chemistry	132	(6·47)	1431	(7·01)	−0·54	3·61	3·46	0·15
Commerce	154	(7·55)	1481	(7·25)	0·30	3·42	3·42	0·00
English	1166	(57·13)	11 607	(56·85)	0·28	3·06	3·04	0·02
French	503	(24·64)	5124	(25·10)	−0·46	3·01	3·02	−0·01
Geography	867	(42·48)	8318	(40·74)	1·74	3·13	3·11	0·02
German	39	(1·91)	351	(1·72)	0·19	2·92	2·92	0·00
History	653	(31·99)	6450	(31·59)	0·40	3·14	3·23	−0·09
Home econ. (Cookery and hostess)	257	(12·59)	2448	(11·99)	0·60	3·14	3·16	−0·02
Integrated science	149	(7·30)	1590	(7·79)	−0·49	3·79	3·73	0·06
Mathematics	1169	(57·28)	11 538	(56·51)	0·77	3·38	3·37	0·01
Metalwork	202	(9·90)	2063	(10·10)	−0·20	3·42	3·43	−0·01
Needlework (Fashion)	117	(5·73)	1247	(6·11)	−0·38	3·26	3·31	−0·05
Physics	354	(17·34)	3588	(17·57)	−0·23	3·37	3·37	0·00
Religious knowledge	197	(9·65)	2105	(10·31)	−0·66	3·67	3·66	0·01
Technical drawing	395	(19·35)	3835	(18·78)	0·57	3·49	3·45	0·04
Woodwork	186	(9·11)	1938	(9·49)	−0·38	3·24	3·27	−0·03

A glance down the column showing the differences in the percentage occurrence between the 1 in 10 sample and the population shows that apart from geography (1·74) only art and crafts approaches a 1 per cent difference. Such differences are small, and are not statistically significant overall.

A comparison of the mean grade figures in the right-hand side of Table VII.2 shows that the values of the differences for biology, history and chemistry (the largest obtained) are about a tenth of a grade or more. Such differences are not statistically significant. In the light of Tables VII.1 and VII.2, therefore, there is little concern for the method of sampling as it may be seen that there is no evidence of bias† in the sample.

Now that the satisfactory relationship between the 1 in 10 sample and the population has been established, it is possible to make a realistic comparison between the three proposed methods for estimating the severity of subjects by using a very large number of candidates (20 416), and to compare the sample and population estimates obtained for each of the internal methods. The ANOVA method was, however, only used on the 1 in 10 sample (in contrast to the subject-pair method and the UBMT method which were used on both the sample and the population), as the computing involved for analysing the population would have been enormous. In order to make any comparisons as pertinent as possible, the ANOVA analysis was performed on the same basis as the other methods, namely that candidates attempting only a single subject were dropped from the analysis. In addition, to check the stability of the estimates within the given sample, the ANOVA analysis was conducted for the candidates remaining when those attempting three or fewer subjects were dropped from the analysis. Table VII.3 presents the results of these analyses (in the ANOVA columns, sample 1 and sample 3 refer to the two analyses mentioned above).

Before discussing the sample and population estimates, the sample 1 and sample 3 results of the ANOVA analyses are worthy of consideration. The mean difference between the pairs of estimates for each subject is only 0·02 of a CSE grade. Only for chemistry, commerce, French and German do the comparisons exceed 0·04 of a CSE grade. A comparison of the numbers of candidates involved in the two analyses and the mean grades obtained is presented in Table VII.4 together with the same information for all candidates in the 1 in 10 sample. The final column of this table gives the mean number of subjects attempted for candidates attempting each subject. It is on this mean number of subjects, minus one, that the UBMT values in Table VII.3 are based.

† It should be noted that the term bias is used here in the statistical sense (i.e. lack of representativeness) rather than in the sense used in the previous chapter (i.e. whether or not a candidate's grade in a subject contributes to the determination of his ability when considering the severity of that subject).

Table VII.3 Population mean grade severities and sample estimates for the CSE board

	METHOD					
	Subject pair		UBMT		ANOVA	
Subject	*Sample*	*Population*	*Sample*	*Population*	*Sample 1*	*Sample 3*
Arithmetic	0·53	0·45	0·57	0·53	0·44	0·39
Art and crafts	−0·18	−0·24	−0·19	−0·23	−0·20	−0·17
Biology	−0·15	−0·06	−0·08	−0·02	−0·07	−0·07
Business studies	−0·28	−0·08	−0·06	−0·03	−0·07	−0·10
Chemistry	0·69	0·32	0·40	0·29	0·44	0·50
Commerce	0·04	−0·02	−0·07	0·01	−0·07	−0·02
English	−0·53	−0·49	−0·56	−0·54	−0·48	−0·49
French	0·17	0·25	0·08	0·17	0·15	0·20
Geography	−0·34	−0·32	−0·32	−0·33	−0·30	−0·29
German	0·59	0·31	0·48	0·29	0·56	0·51
History	−0·30	−0·24	−0·27	−0·22	−0·26	−0·25
Home econom. (Cookery and hostess)	−0·13	−0·22	−0·31	−0·31	−0·30	−0·30
Integrated science	0·17	0·11	0·18	0·12	0·13	0·13
Mathematics	0·13	0·17	0·16	0·18	0·10	0·07
Metalwork	−0·07	−0·06	−0·10	−0·04	−0·08	−0·10
Needlework (Fashion)	−0·24	−0·25	−0·02	−0·10	−0·04	−0·04
Physics	0·30	0·18	0·10	0·11	0·10	0·09
Religious knowledge	0·38	0·46	0·36	0·32	0·32	0·35
Technical drawing	−0·23	0·03	−0·01	0·04	−0·02	−0·02
Woodwork	−0·44	−0·24	−0·35	−0·25	−0·34	−0·38

Table VII.4 shows that while the numbers of candidates fall from the basic 1 in 10 sample to the ANOVA 1 and ANOVA 3 samples, the mean grade stays reasonably constant, as might be expected. Two subjects stand out, however, as being very different: French and German. It would appear that, of the 503 candidates attempting French, only 395 attempted at least one other subject; the difference between these numbers, 108 (22 per cent), represents those candidates attempting *only* French. Unfortunately the sample figures for German are too small to allow for the corresponding calculation to be meaningful. Similar figures for French and German in the population are given in Table VII.5.

Thus, about a quarter of those taking French and German in the CSE board

take only these subjects and, by looking at the fall in the mean grade that occurs when they are removed, it may be seen that they are better candidates than the group as a whole taking these two subjects. The importance of these figures to the estimation of the severity of French and German is that the estimates for these two subjects may not be a true reflection of that obtained when candidates attempting a single subject are included in the analysis. In addition, since they attempt overall a low number of subjects (4·25 and 3·29 respectively, see Table

Table VII.4 A comparison of the ANOVA samples 1 and 3 with the 1 in 10 sample in terms of number of candidates, the mean grade and the mean number of subjects taken in the CSE board

	NO. OF CANDIDATES			SUBJECT MEAN GRADE			*Mean no. of subjects attempted* ANOVA *sample 1*
Subject	*1 in 10 sample*	ANOVA *sample 1*	ANOVA *sample 3*	*1 in 10 sample*	ANOVA *sample 1*	ANOVA *sample 3*	
Arithmetic	239	225	172	4·00	4·02	4·01	4·53
Art and crafts	404	385	326	3·40	3·45	3·44	5·14
Biology	232	230	186	3·21	3·21	3·20	4·98
Business studies	110	109	84	3·39	3·39	3·41	4·57
Chemistry	132	125	101	3·61	3·64	3·77	5·28
Commerce	154	149	125	3·42	3·44	3·43	4·70
English	1166	1138	932	3·06	3·05	3·02	4·91
French	503	395	233	3·01	3·15	3·30	4·25
Geography	867	847	711	3·13	3·15	3·16	5·04
German	39	28	11	2·92	3·14	3·36	3·29
History	653	640	563	3·14	3·15	3·18	5·22
Home economics (Cookery and hostess)	257	250	201	3·14	3·18	3·18	4·79
Integrated science	149	143	126	3·79	3·78	3·81	5·23
Mathematics	1169	1044	778	3·38	3·44	3·48	4·70
Metalwork	202	198	181	3·42	3·42	3·43	5·69
Needlework (Fashion)	117	112	99	3·26	3·31	3·26	5·14
Physics	354	348	289	3·37	3·39	3·44	5·28
Religious knowledge	197	192	166	3·67	3·72	3·74	5·20
Technical drawing	395	385	342	3·49	3·48	3·49	5·43
Woodwork	186	180	159	3·24	3·28	3·26	5·33
No. in sample	2041	1626	1092	—	—	—	—

Table VII.5 Analysis of candidates attempting French and German in the CSE board population

	NO. OF CANDIDATES			SUBJECT MEAN GRADE		
	All population	*Those attempting two or more subjects*	*Difference*	*All population*	*Those attempting two or more subjects*	*Difference*
French	5124	4018	1106 (22%)	3·02	3·17	−0·15
German	351	254	97 (28%)	2·92	2·97	−0·05

VII.4) the value obtained is likely to be rather more unreliable than that obtained for many of the subjects.

The comparisons of the estimates of the severity of the twenty subjects using the subject-pair and UBMT methods for the sample and population as presented in Table VII.3 are now summarized in a similar manner to that found in Chapter VI. Table VII.6a uses ANOVA sample 1 as a basis for comparison while Table VII.6b presents the comparisons for the subject-pair and UBMT methods only.

Comparing with the ANOVA 1 values first of all, it is clear that the UBMT sample values agree very well (0·03) over all subjects. The subject-pair values are much more widely spread (0·10). The population values for both methods

Table VII.6 An empirical comparison of the subject-pair, UBMT and ANOVA estimates of severity with each other and their population values

a Comparison with ANOVA 1

Method and group	*Mean difference from* ANOVA 1 (*CSE* grades)
Subject-pair sample	0·10
Subject-pair population	0·07
UBMT sample	0·03
UBMT population	0·06
ANOVA 3	0·02

b Subject-pair and UBMT comparisons mean differences (CSE grades)

Method and group	*Subject-pair population*	UBMT *sample*
Subject-pair sample	0·11	0·10
UBMT population	0·05	0·06

agree reasonably well (0·06 and 0·07) with the ANOVA 1 sample. From Table VII.6b, a reasonable degree of agreement is seen in the bottom row (UBMT population) as compared with the top row (subject-pair sample). The main reasons for these differences are now summarized.

First, as has been seen above, the subject-pair method is susceptible to small numbers of candidates attempting any pair of subjects. It is likely that this fact causes the lack of agreement between the subject-pair sample estimates and any other set of estimates of severity. The subject-pair population values agree fairly well with other sets of values, and this may be taken as a suggestion that the problem of sample sizes is not insuperable – given enough candidates the method may be used. Even if the population were small, there is no problem about using the method if its philosophy is accepted.

Secondly, the UBMT sample and the UBMT population agreement (and similarly the subject-pair sample and population agreement) is marred by the occurrence of relatively large discrepancies in French, German and woodwork; the discrepancy is very likely to be due to sampling error in the case of German, and somewhat less likely in the case of French. With these subjects removed, the mean agreement falls from 0·06 to 0·04 CSE grades. It is a failing of these two internal methods to be sensitive to the situation as discovered with candidates attempting French and German.

As far as minority subjects are concerned, it is clear that, unless peculiar entry patterns are found (as in French and German) there are not likely, on the evidence produced for the CSE board, to be any large deviations from population values in estimates of severity found in any of the methods in a sample of candidates, with the possible exception of the subject-pair method. Even with this method, however, the differences are not great. Comparing the UBMT sample and population values gives a mean difference of only 0·06 CSE grades (about one per cent of the grade scale), so there is little justification in looking at the population values (as opposed to the sample estimates) even for minority subjects.

The main aims of accepting the population of CSE candidates' results for the CSE board have been achieved: in this single exercise the '1 in N' method of sampling has been shown to be perfectly acceptable, the stability of the various methods has been seen to depend more on the entry patterns than on the sample or method used; and it has also been seen that minority subjects may be investigated in a sample, always provided that the peculiar entry patterns are not present.

Finally, there are population results for the CSE sector which again show up the consistency with which the three internal methods, based on different assumptions and philosophies, produce results that are effectively equivalent. With the estimates themselves varying from half a CSE grade severe to half a CSE grade lenient (an effective range of 17–20 per cent of the CSE grade scale)

the differences between methods of 0·05 or so of a CSE grade are themselves negligible.

Use will be made of the results obtained above in Chapter VIII, but it is only fitting that, within the context of this chapter, the help and co-operation of the CSE board should be acknowledged. Through them much has been learned about the methods proposed, and the sampling techniques employed here and elsewhere.

VIII. Some comparisons of results

The greater part of this report concentrates upon the methods by which comparability between subjects may be investigated, and the philosophical and technical problems involved in such investigations. This emphasis is desirable because the data relate to the examinations of 1968 and because the data may have come from unrepresentative samples. Nevertheless the results of the analyses are of some interest in their own right; comparisons of the results obtained in different boards and in different years are particularly important, since any consistency of patterns in the results would lend important support to the validity of the methods of investigation. This concept of replication of results has always been stressed as being very important in the context of comparability research (see, for example, Schools Council Working Paper 34[1]), as indeed it is in any scientific enquiry.

The comparisons to be drawn are fourfold: between GCE boards, between CSE boards, between the GCE and the CSE sectors, and between years in the one GCE board where such a comparison is possible. Results from the ANOVA method are used wherever possible for the sake of uniformity and simplicity.

Comparisons between GCE boards

The estimates of mean grade severity in the ten major subjects studied in Chapters II to VI for GCE boards 1 to 4 are shown in Table VIII.1. The figures in brackets provide the rank order of the severities within each board.

Comparisons of the estimates of severity or leniency for a given subject across the four boards should be made with caution since the estimates for each board are calculated in relation to the consensus within that board. It is more appropriate to compare the rank orders of the subjects in each board and it is apparent that there is a fairly high degree of consistency in the rank orders. Biology moves by five places in the order, and mathematics and physics by three places, but the remaining subjects vary by two places or less.

Comparisons between CSE boards

In the 1968 CSE Monitoring Experiment, from which the data used in the present investigations were derived, the numbers of candidates sampled to represent each CSE board were rather smaller than the numbers in GCE

Table VIII.1 Rank orders and estimates of severity for GCE boards 1–4 (ANOVA method)

Subject	*GCE board* 1	2	3	4
Art	−0·68 (9)	−0·69 (10)	−0·63 (9)	−0·78 (10)
Biology	0·58 (2)	−0·04 (7)	0·23 (4)	0·10 (5)
Chemistry	0·83 (1)	0·53 (1)	0·71 (1)	1·03 (1)
English language	−0·95 (10)	−0·64 (9)	−0·67 (10)	−0·74 (9)
English literature	−0·28 (8)	−0·29 (8)	−0·04 (6)	0·01 (6)
French	0·27 (4)	0·43 (3)	0·63 (2)	0·47 (2)
Geography	−0·20 (7)	−0·01 (6)	−0·54 (8)	−0·25 (7)
History	0·17 (5)	0·21 (4)	0·50 (3)	0·18 (4)
Mathematics	−0·09 (6)	0·04 (5)	−0·40 (7)	−0·47 (8)
Physics	0·36 (3)	0·46 (2)	0·22 (5)	0·46 (3)

boards 1 to 4. (Details of the CSE board sample sizes, mean grades, etc. can be found in Schools Council Working Paper 34.[1]) The CSE board samples were therefore less suitable for methodological research, but since there is evidence supporting their representativeness it is appropriate to give examples of the results obtained using the ANOVA method, which appears most suitable for work with small samples. Table VIII.2 gives the estimates of severity and their rank orders for three CSE boards, which were chosen on the grounds that their samples were larger than those of the other CSE boards.

Table VIII.2 Rank orders and estimates of severity for CSE boards 18, 21 and 24 (ANOVA method)

Subject	*CSE board* 18	21	24
Art	−1·02 (9)	−0·94 (9)	−0·66 (9)
Biology	0·13 (5)	−0·18 (6)	0·01 (5)
Chemistry	0·25 (3)	0·37 (3)	0·63 (1)
English	−0·63 (8)	−0·56 (8)	−0·44 (8)
French	0·76 (1)	0·70 (2)	0·58 (2)
Geography	−0·28 (7)	0·06 (4)	−0·04 (6)
History	0·00 (6)	−0·21 (7)	−0·30 (7)
Mathematics	0·57 (2)	0·77 (1)	0·12 (3)
Physics	0·22 (4)	−0·01 (5)	0·11 (4)
No. of candidates in sample	882	824	694

As with the four GCE boards, the rank orders of subjects are fairly consistent across the three CSE boards. Geography moves three places in the order of ranking and mathematics and chemistry two places, while the remaining subjects vary by only one place or, in two cases, not at all.

Comparisons between the GCE sector and the CSE sector

A direct comparison of the estimates of severity between GCE and CSE boards is not appropriate, as the grade scales are not the same. It is possible, however, to compare the orders of ranking of the subjects in the two sectors. Table VIII.3 presents these results (English literature has been dropped from the results for the GCE sector since there is no equivalent subject in the CSE sector).

Table VIII.3 Comparison of rank orders of severity estimates in the GCE sector and the CSE sector (ANOVA method)

	GCE board				*CSE board*		
Subject	1	2	3	4	18	21	24
Art	8	9	8	9	9	9	9
Biology	2	7	4	5	5	6	5
Chemistry	1	1	1	1	3	3	1
English	9	8	9	8	8	8	8
French	4	3	2	2	1	2	2
Geography	7	6	7	6	7	4	6
History	5	4	3	4	6	7	7
Mathematics	6	5	6	7	2	1	3
Physics	3	2	5	3	4	5	4

It is apparent that there is a higher degree of consistency within each sector than there is between the two sectors. Nevertheless, art and English consistently appear as the two most lenient subjects, while chemistry and French tend to be the most severe. The most noticeable differences between the two sectors occur in history and mathematics.

Comparisons between years

A comparison of results between years is of interest only within any particular board. Data are available in the case of one GCE board, the Joint Matriculation Board, who have investigated comparability between subjects using Aptitude Test 100 in the years 1970 and 1971.[2,3] Their results may be compared with those for 1968 as reported in Chapter II, using the regression method with Test 100.

It is important to note that the 1968 NFER study differed from the two JMB studies in the following three aspects, all likely to be of some methodological importance.

a In the 1968 study, it was impossible to distinguish between different syllabuses or alternatives within a subject; English language, for example, represents an unknown mixture of the JMB alternative English language schemes.

b The 1968 samples have not been shown to be representative of the 1968 JMB population; with a few isolated exceptions, the 1970 and 1971 samples have been shown to be representative of their respective populations.

c The 1970 and 1971 JMB samples for each subject are independent, in that candidates in the sample for geography, for example, do not appear in the sample for any other subject; in contrast, the 1968 samples for each subject overlap to a considerable extent (for example, a candidate in the geography sample may also appear in the samples for several other subjects).

Table VIII.4 presents the results of the 1968, 1970 and 1971 studies for the JMB. In the JMB 1971 study, subjects other than those shown were included in the analyses and revealed, on average, a tendency towards leniency. As a consequence the 1971 estimates of severity for the subjects shown in Table VIII. 4 do not sum to zero; this means that the values of the severities for each subject cannot be compared directly, but the rank orders of the severities are still of interest. The rank orders are, of course, unaffected by the location of the zero point of the scale.

There are a number of changes in the rank orders across the years, with mathematics showing the most marked variation. Nevertheless, the directions of the deviations from zero show considerable consistency and English language and English literature invariably appear at the bottom of the rank order. There is as good agreement between the 1968 results and those of the other two years as there is between the results of the two JMB studies, which lends support to the idea that the 1968 sample is likely to be representative of the 1968 population. This consistency also lends support to the use of overlapping samples. In the 1970 JMB study some 9000 candidates were tested, with approximately 1000 candidates in each of the two English language samples, while in the 1968 NFER study some 3400 candidates were tested with approximately 2800 candidates in the sample for English language. Although in theory the NFER procedure (which was employed for collecting data in another type of investigation) is not as good as the JMB procedure in the context of subject comparability, in practice its results appear equally satisfactory and its economic advantages are obvious. The research reported in Chapter IX, however, leads to the conclusion that the proportions of each sex in the sample must be representative

Table VIII.4 Rank orders and estimates of severity for JMB in 1968, 1970 and 1971 (regression method)

Subject	*NFER 1968*	*JMB 1970*	*JMB 1971*
Art	−0·49 (9·5)	−0·25 (8)	
Biology	−0·14 (7)	−0·16 (7)	0·36 (5)
Chemistry	0·33 (2)	0·64 (2)	0·32 (6)
English language [a]	−0·49 (9·5)		
Paper A		−0·46 } (9·5)	−0·37]
Paper B		−0·61]	−0·80 } (9)
School assessment			−0·75]
English literature [a]	−0·24 (8)		
Paper A		−0·56 (9·5)	0·02] (8)
Paper B			0·65]
French	0·25 (3)	0·30 (4)	0·61 (1)
Geography	0·09 (6)	0·01 (5)	0·43 (2)
History	0·16 (4)	−0·15 (6)	0·37 (4)
Mathematics	0·12 (5)	0·31 (3)	0·15 (7)
Physics	0·37 (1)	0·66 (1)	0·38 (3)

[a] The ranks for English language and English literature in 1970 and 1971 are rough figures based on a consideration of the relative popularity of the alternative schemes.

of the population proportions; this may be difficult to achieve in every subject with the NFER sampling procedure.

Severity and the calibre of candidates

It is of interest to examine the relation between the severity of a subject and the calibre of candidates offering it. This may be done by plotting, for each subject, the severity against the mean ability of the candidates offering that subject. The graph can be drawn for any of the methods described in this report but here it has been done for the ANOVA method. The mean ability is plotted along the vertical axis and the severity along the horizontal axis. However it must be remembered that the ANOVA estimate of a candidate's ability $(a + c_i)$ (see Chapter V, p. 39) is small for high abilities and large for low abilities, since it is based on the grades obtained by a candidate, and the standard of performance decreases with the magnitude of the grade awarded. The reader is also reminded that allowance has already been made for differences in the calibre of individual candidates and Figure 5 should be interpreted with this in mind.

The figure shows the graph plotted for all candidates of board 2 and there is

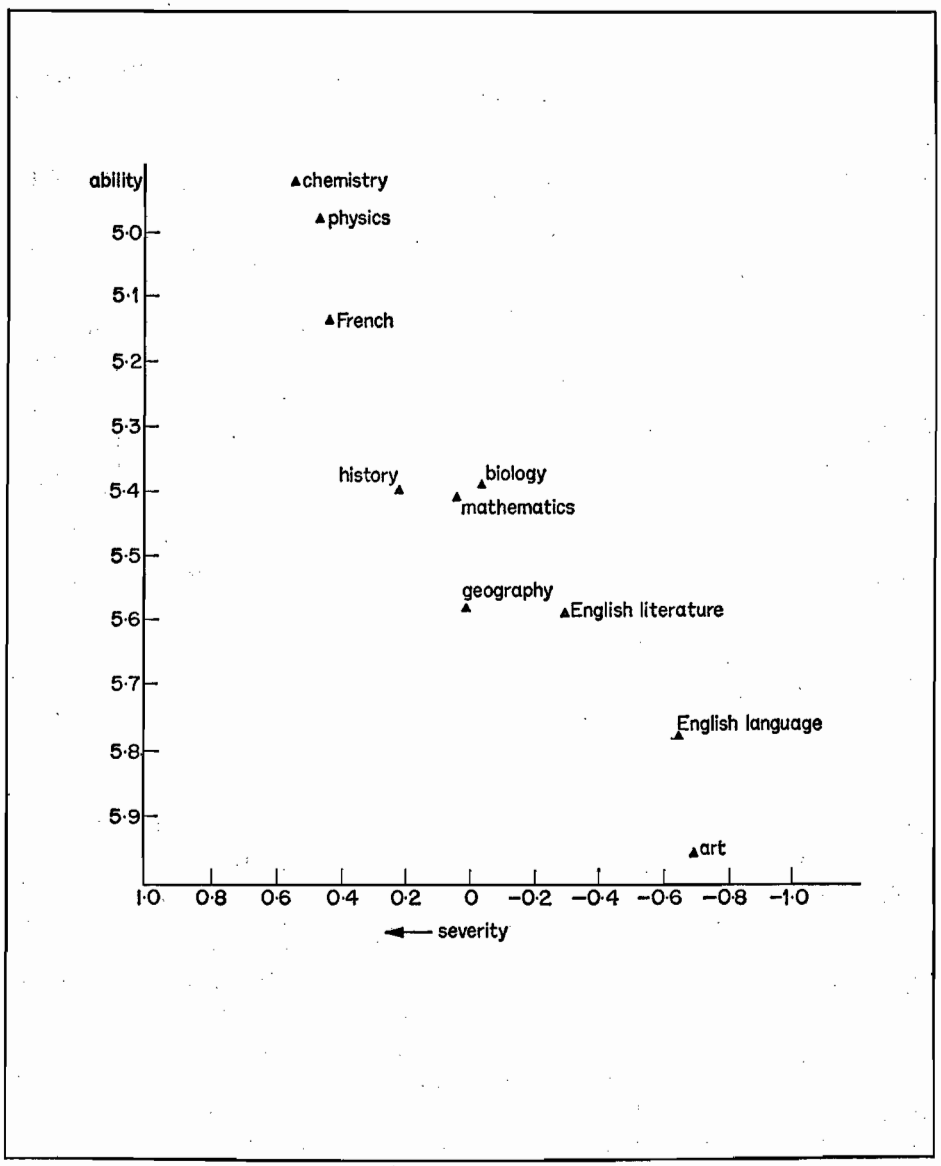

Fig. 5 *Plot of ability against severity for GCE board 2* (ANOVA *method*)

clearly a tendency for high severities to be associated with high abilities (for example, chemistry, physics and French), and for low severities to be associated with low abilities (for example, English language and art). Table C.4 (p. 107) will enable the reader to plot graphs for the three other GCE boards and a similar tendency will be noticed.

Discussion

The consistency of the results between boards, sectors and years is strongly suggestive of real differences between subjects rather than of differences which have arisen because of sampling. This conclusion is given further support by the results reported in Chapter VII for the population of candidates in one CSE board, which reveal a very similar pattern of differences between subjects.

Given that the differences are real, the consistency of the pattern is not perhaps surprising. The GCE and CSE boards go to great lengths to ensure comparability of standards in each subject between themselves, through cross-moderation studies and discussion of statistical evidence, and the 1968 CSE Monitoring Experiment revealed no evidence of a lack of comparability of standards in each of a wide range of subjects.[1] The consistency of the pattern of the results between the GCE sector and the CSE sector is also to be expected, by virtue of the accepted comparability of standards between a pass at GCE O level and CSE grade 1 in any given subject. Indeed, it could be argued that the CSE boards had no choice but to generate different standards between subjects if they were to secure the comparability of grade 1 with a pass at O level in every subject. Perhaps the greatest effort towards comparability of standards is made within each board in the area of comparability between years, so the consistency of the pattern of differences between subjects that emerged in the JMB in the years 1968, 1970 and 1971 is perhaps the least surprising of all the comparisons.

On the basis of the results presented in this report, it seems possible to conclude that English language and English literature are consistently lenient. It seems probable that art is also lenient, but, since the skills and abilities required in art are rather different from those required in the other nine predominantly academic subjects to which most attention has been paid in this report, the case is less conclusive. At the other end of the scale, French and chemistry are consistently severe in both sectors, together with physics in the GCE sector and mathematics in the CSE sector.

The terms 'lenient' and 'severe' have been used to indicate the direction of the differences from the consensus standard. In interpreting the results, it must be made clear that the differences may be due to factors other than differences in grading standards adopted by examiners in different subjects, and hence that

the terms 'lenient' and 'severe' are a convenient shorthand. The possible importance of candidates' motivation has already been mentioned and it cannot be denied that part of the observed differences between subjects may be explained by differences between groups of candidates entering for different subjects that cannot be detected by any of the techniques used (the effect of school type, for example).

The picture that emerges accords fairly well with common suppositions about the conceptual difficulty of subjects. Perhaps the outstanding example is that of Latin, which was found to be the most severely graded of twelve subjects in board 2 (Table V.1, p. 40); this is generally regarded as a difficult subject and there is a tendency for schools to offer it only to their more able pupils. French is perceived by pupils as being one of the more difficult subjects they study (see *French in the Primary School: Attitudes and Achievement* by Clare Burstall[4]). This conjecture is lent support by the evidence itself: French and the physical sciences tend to be taken by more able pupils. However, all this is not to argue that the subjects that appear severe are necessarily of greater conceptual difficulty. Nevertheless, it appears that the more severely graded subjects tend to be offered by candidates of above average ability.

It has been suggested to the writers that since the aims, content and teaching of subjects differ, it is inappropriate to compare performance in different subjects. In other words, comparability of standards between subjects is a red herring and any apparent differences in standards between subjects that emerge in such analyses have in fact nothing to say about standards but only something to say about subjects.

We would be happy to accept these arguments for an examination system which was criterion-referenced, where, for example, to obtain a pass in English language a candidate had to demonstrate his ability in a defined set of skills. The examination system partly functions like this, but it is also a norm-referenced system, most obviously in the CSE sector (as argued in Chapter I). Whatever the particular skills acquired by a 16-year-old in a particular subject, he may expect to obtain a grade 4 if his attainment is average when compared with his peers'. While standards in GCE O level are not defined in the same way, they are taken to be so operationally by those users who demand a specified number of passes without defining all of the subjects in which these passes must be obtained, implying a willingness to consider a pass in history as equivalent to one in geography. The GCE boards themselves recognize that the percentage passing the examination should vary with the calibre of the candidates, as the following quotation shows:

> It is a myth that the Ordinary level pass standard is determined as a percentage of the entry. It may seem so, since in mass entry subjects with

> candidates of much the same average calibre from one year to the next, it is unlikely that the proportion passing or failing to reach a fixed standard of judgment will vary appreciably over short periods of time. But even the most cursory examination of the figures for subjects with variable calibre entries (typically, 'new' subjects with a small initial entry) shows that variability of calibre is reflected in the results. Indeed, such variability has sometimes been a source of misdirected criticism of the work of the boards.[5]

We acknowledge that the problem may not be as great as it might appear on the surface. It is probable that in practice relatively few users equate a pass in English language with a pass in French. Where they do 'trade' a pass in one subject for a pass in another, it is more likely to be within a group of subjects that have some affinity (for example, one modern language for another, one science for another) but this is by no means always the case. The results suggest that differences between subjects within such groups are smaller than those between groups. We would certainly accept that subject comparability is of more importance in practice than in theory and, if users confined their 'trading' of passes to within groups of similar subjects, comparability between subjects might well become less of a problem. At least techniques are now available for investigating at least the extent of differences between subjects, whether across the full range or within groups of like subjects.

There is, however, one complicating factor in the investigation of subject comparability (on whatever scale it is undertaken) which has been ignored until now in this report. It concerns the differences in examination performance in different subjects between the sexes. Chapter IX reveals and discusses these differences and Chapter X relates them to the broader issues.

References

1. D. L. NUTTALL, *The 1968 CSE Monitoring Experiment* (Schools Council Working Paper 34). Evans/Methuen Educational, 1971.
2. G. M. FORREST, *Standards in Subjects at the Ordinary Level of the GCE, June 1970* (Occasional Publication 33). Joint Matriculation Board, Manchester, 1971.
3. G. M. FORREST and G. A. SMITH, *Standards in Subjects at the Ordinary Level of the GCE, June 1971* (Occasional Publication 34). Joint Matriculation Board, Manchester, 1972.
4. CLARE BURSTALL, *French in the Primary School: Attitudes and Achievement.* NFER, Slough, 1970.
5. Schools Council, *Examining at 16+: the Report of the joint GCE/CSE Committee of the Schools Council.* HMSO, 1966.

IX. The question of sex differences

The problem of concern in this chapter is ostensibly a simple one: 'What sex differences are observed in performance in examinations at 16+?' At the simplest level a straight answer may be given by presenting a table of mean grades of boys and girls in different subjects. Thus Table IX.1 gives the mean grades by sex obtained in board 2, and further tables will be found in Appendix B (Tables B.10, 11, and 14–17).

It is also possible to report across the boards in order to show the overall pattern of relative achievement between the sexes. Table IX.2 indicates whether the boys or the girls have obtained the better mean grade for each subject and each board. Thus in geography in board 2 it may be seen from Table IX.1 that the boys had the better (numerically lower) mean grade and this is shown by placing 2 opposite 'geography' in the section under 'boys'.

Having included Table IX.2 it is necessary to caution the reader about the crudity of its approach. First of all, some of the pairs of mean grades are very close, as in French in board 2 where the difference is 0·01 or in history in the same board with a difference of 0·10. Small differences are clearly liable to go the other way when another sample is taken, so there might be a number of changes in the table had different candidates been selected for the samples. The second objection is that there is no guarantee that the samples in any of the boards are representative of the board as a whole, with the same implications as

Table IX.1 Sample mean grades by sex for GCE board 2

Subject	*Boys*	*Girls*
Art	5·29	5·27
Biology	4·97	5·48
Chemistry	5·17	5·87
English language	5·33	5·02
English literature	5·99	4·97
French	5·94	5·35
Geography	5·10	5·91
History	5·55	5·64
Mathematics	5·10	5·75
Physics	5·51	5·23

Table IX.2 Better mean grade of the sexes by subject across the GCE boards

Subject	*Boys*								*Girls*							
Art			3					8	1	2		4	5	6	7	
Biology		2	3		5			8	1			4		6	7	
Chemistry		2	3	4	5			8	1					6[a]	7	
English language									1	2	3	4	5	6	7	8
English literature									1	2	3	4	5	6	7	8
French						6		8	1	2	3	4	5		7	
Geography	1	2	3		5	6		8				4			7	
History		2							1		3	4	5	6	7	8
Mathematics		2	3	4	5		7	8	1					6		
Physics				4	5		7	8	1	2	3			6[a]		

[a] only one candidate

above. Thirdly, the analysis takes no account of the calibre of the candidates in the two sexes. It could be argued, for instance, that the more able girls are interested in arts subjects and the more able boys in science subjects. Further, in the case of subjects that are offered by schools as options, it is not known whether the proportions of able pupils in the two sexes are comparable. For instance, it might be the case that art is commonly regarded with approval in girls' schools and thought of as a soft option in boys' schools. So it becomes necessary to consider the problem of sex differences in a less unsophisticated way than might have been the case had answers to some of the above points been known.

Test 100 as a measure of calibre

It was suggested in the last paragraph that Tables IX.1 and IX.2 take no account of the calibre of candidates entering for the different subjects. However, the scores on Aptitude Test 100 referred to in Chapter I are available. Table IX.3 gives the mean scores on this test by sex and board. It will be seen that in every case the mean score of boys on Test 100 is better than that of the girls in the same board. However, examining the overall mean grade, irrespective of subject, for boys and girls in the different boards (see Table IX.4) shows clearly that girls in these samples tend to obtain better grades at GCE O level than boys. (In the board 8 sample there are only 94 girls.)

The superiority of boys judged by Test 100, and the superiority of girls judged by GCE grade, raises the question of whether Test 100 is a suitable measure of the calibre of the candidates when making comparisons between the

Table IX.3 Sample mean Test 100 scores by sex and GCE board

GCE board	*Boys*	*Girls*
1	49·4	48·3
2	52·3	46·8
3	50·8	46·8
4	55·1	48·6
5	50·8	44·9
6	54·7	47·1
7	45·1	38·4
8	50·7	40·5

sexes. One major difficulty is that it is hard to know precisely what this 'calibre' consists of, and it is by no means obvious that there is no sex bias to it. Consider the case of a physiologist who is interested in sex differences in the strength of different muscles. Here there should be no hesitation in saying that men tend to be stronger than women, and, having admitted this, it would be possible to go on to investigate sex differences in tests of the strength of various muscles. It may be objected that in the cognitive domain there is little evidence of the overall superiority of either sex. While accepting this, it should be added that success in GCE O level draws on traits of personality, and immediately we are on very uncertain ground: first, there are personality differences between the sexes and secondly, it has not been established what particular traits are important in passing the examination, nor whether there is a sex bias in any of them. The problem would not be solved either if representative samples of the candidates in the different boards were available, because it is not known if there is some sex discrimination acting through our society which makes the calibre of the boy and girl candidates not quite comparable.

Table IX.4 Sample mean grades over ten subjects by sex and GCE board

GCE board	*Boys*	*Girls*
1	5·60	4·72
2	5·41	5·38
3	5·41	5·18
4	5·62	5·30
5	5·57	5·17
6	5·26	4·69
7	6·30	5·96
8	6·00	6·42

While Nuttall[1] and Forrest[2] came to the conclusion that these scores were suitable for their purposes of comparing boards and subjects within a board, the reversal of the sex superiority between Table IX.3 and IX.4 leads to the conclusion that scores on Test 100 are not suitable for making comparisons between the sexes. Accordingly, it becomes necessary to look elsewhere for an alternative measure of the calibre of candidates.

c_i as a measure of calibre

In Chapter V it was suggested that the statistic c_i provided a measure of the overall ability of candidates. The argument in favour of this may be put as follows: 'It must be assumed that examiners mark their papers fairly and that knowledge of the sex of the candidate does not affect the grade awarded. Even

Table IX.5 Mean correlation of c_i and of Test 100 score with GCE O-level grade by GCE board

GCE board	c_i	*Test 100*[a]
1	0·737	0·353
2	0·710	0·357
3	0·729	0·354
4	0·716	0·356

[a] These are overall values; those given in Appendix B, Tables B.5 and B.10–17 are pooled within-school values.

if there is a tendency for girls to offer subjects that are more leniently graded, differences in the overall severity of marking of subjects are taken into consideration in the calculation of c_i. In general a number of subjects contribute to the c_i of candidate i, so that clearly it will be correlated with ability to perform well in GCE O level.' This last aspect can be measured and compared with the correlation of scores on Test 100 with subject grades; c_i clearly has the advantage as a measure of ability to succeed in O level (see Table IX.5).

While there is at least a prima facie case for taking c_i to be a measure of the calibre of the candidates, the statistic suffers from the disadvantage that it is not an independent estimate and, because of the tendency of the girls in the samples to do better in O level than the boys, is suspect of being biased in favour of girls. Whether this is a fair criticism is open to doubt – it may well be appropriate in the present circumstances to take a measure of calibre which favours girls. For, if the calibre is to be ability to do well in GCE O level, and girls do

better than boys in the examination, then it is appropriate that our measure of calibre should be biased in favour of girls. This argument appears to carry still more weight when the reader is reminded that the purpose of the measure of calibre is simply to compare grades obtained by boys and girls *within* the examination.

Estimates of sex difference in subjects

It is now possible to describe and compare four methods of estimating grade differences for the sexes in the various subjects:

i The mean grade for each sex is found and the difference taken:

(mean grade of girls) minus (mean grade of boys).

The obvious disadvantage of this method is that it makes no allowance for the possibility that the calibres of the boys and girls offering a particular subject may be different. For this reason it will be termed the 'uncontrolled difference'.

ii The severity for each sex is found and the difference taken:

(severity for girls) minus (severity for boys).

This method takes account of the calibre of candidates in a particular subject (in relation to the calibre of candidates of the same sex in the sample) but assumes that the scales on which severity is measured for boys and girls have the same starting-point or origin.

iii A statistical analysis is carried out to estimate the difference in subject grades between the sexes when allowance has been made for difference in the calibre of individuals as measured by c_i (calculated from an analysis of variance carried out on the sample as a whole). The method used is an analysis of co-variance. The disadvantage of this method is that c_i may be biased in favour of girls but the bias is believed to be less than that of the fourth method.

iv The same type of statistical analysis as in **iii** is made for differences in the calibre of individuals as measured by Test 100. The disadvantage of this method is that Test 100 is believed to be considerably biased in favour of boys.

The results of applying these methods to board 2 are shown in Table IX.6, and to boards 1, 3 and 4 in Appendix C (Tables C.5–7). In all four cases a positive value means a difference in favour of boys and a negative value indicates a difference in favour of girls.

The columns have been arranged to bring out two features. The first of these

Table IX.6 Grade differences between the sexes by subject in GCE board 2 using four different methods

Subject	METHOD OF ESTIMATING GRADE DIFFERENCE			
	Difference between severities (**ii**)	*Allowance made for* c_i (**iii**)	*Uncontrolled difference* (**i**)	*Allowance made for Test 100* (**iv**)
Art	0·18	0·21	−0·02	−0·14
Biology	0·28	0·35	0·51	−0·08
Chemistry	0·76	0·79	0·70	0·42
English language	−0·53	−0·34	−0·32	−0·92
English literature	−1·40	−1·15	−1·02	−1·55
French	−0·94	−0·75	−0·59	−1·18
Geography	0·82	0·86	0·81	0·32
History	−0·18	−0·05	0·10	−0·31
Mathematics	0·85	0·91	0·65	0·30
Physics	0·18	0·31	−0·28	−0·39

is that, in general, the figures under **iii**, **i**, **iv**, for a given subject in a given board, are decreasing from left to right. It was explained above that there may be a slight bias in favour of girls in c_i, and that there is believed to be a considerable bias in favour of boys in scores on Test 100. This means that entries under **iii**, which make allowance for c_i, can be expected to be larger (more positive) than entries under **iv**, which make allowance for Test 100; this is observed to be so in every case. Exceptions to the general rule that the figures under **iii**, **i**, **iv** are decreasing from left to right are found as follows:

a Entries under **i** and **iv** are reversed compared with the usual order in the case of:

board 1 – physics, art, mathematics and chemistry;
board 3 – physics and art.

In all these cases both c_i and Test 100 indicate that the girls have higher calibre than the boys.

b Entries under **iii** and **i** are reversed compared with the usual order in the case of:

board 2 – history, French, biology, English language and English literature;
board 3 – French and biology.

In all cases both c_i and Test 100 indicate that the boys have higher calibre than the girls.

Table IX.6 (and Tables C.5–7) are also arranged so that a comparison can

easily be made between the entries under **ii** and **iii** (difference in severities and allowance made for c_i). It will be seen that there is quite good agreement between the figures obtained. While there are some discrepancies in the values obtained, the rank orders agree well, as will be seen below.

Interpretation of results

Table IX.6 (and Tables C.5–7, pp. 108–9) show clearly that the choice of the control can make a great difference to the results obtained. It is suggested that the figures under **ii** and **iii** are the best available estimates of sex differences at GCE O level to date, but it is recognized that some may wish to question this suggestion. If, however, scores on Test 100 are considered to be an appropriate measure of calibre, the consequences of this choice are quite dramatic; 77·5 per cent of the differences show that girls are doing better than expected from their Test 100 scores, and 32·5 per cent of the differences are greater than one grade. If the reader prefers to follow the consequences of taking c_i as a measure of calibre, girls did better than might be expected from their values of c_i in 47·5 per cent of the cases examined, and 15·0 per cent of the differences (both positive and negative) are greater than one grade. It is suggested that the best answers (if there are such) lie fairly near to those under **ii** and **iii** of Table IX.6 but, whatever opinion is held as to the choice of a control, the existence of sex differences in the samples is unmistakably shown. For example, girls did better than boys at French, English language and English literature when their estimated calibre is taken into account.

In Table IX.7 the sex difference of Table IX.6 (and Tables C.5–7, pp. 108–9) have been ranked for both controls and for differences in severities in each board. The differences are ranked from 1 (the subject most biased towards girls) through to 10 (the least biased towards girls), and vice versa in terms of bias towards boys. While arranged for comparisons between boards, it is worthwhile comparing the ranks in each board. Here there is quite good agreement but some ranks differ by as much as three.

When comparing figures for sex differences obtained from different boards there is less reason to suppose agreement since different sets of candidates are examined by different means. Nevertheless, it is clear that girls tended to obtain better grades than boys in English language and literature and worse grades than boys in mathematics when their estimated calibre is taken into account. After that the picture is less clear but girls tended to do better than boys in French and, with certain exceptions, worse than boys in geography, physics and chemistry, again when their estimated calibre is taken into account. This is very much in line with previous reports on attainment tests mentioned by F. T. Tyler in the *Encyclopedia of Educational Research*.[3]

Table IX.7 Rank orders of grade differences between the sexes by GCE board and method

Subject	GCE BOARD											
	1	2	3	4	1	2	3	4	1	2	3	4
	Allowance made for Test 100				*Allowance made for c_i*				*Differences in severities*			
Art	10	6	10	4	7	5	7	3	8	5	9	3
Biology	4	7	6	6	4	7	4	6	4	7	4	6
Chemistry	3	10	7	8	6	8	9	8	5	8	8	8
English language	1	3	2	1	2	3	2	1	2	3	2	1
English literature	2	1	1	3	1	1	1	2	1	1	1	2
French	5	2	3	2	3	2	3	4	3	2	3	4
Geography	9	9	8	5	10	9	8	5	9	9	7	5
History	6	5	4	7	5	4	5	7	6	4	5	7
Mathematics	8	8	9	10	9	10	10	9	10	10	10	9
Physics	7	4	5	9	8	6	6	10	7	6	6	10

Further light is shed on the situation by the scores on the two parts of Test 100, which may be described as verbal and quantitative. The difference (verbal score minus quantitative score) may be taken as a measure of verbal/quantitative bias, a positive score indicating verbal and a negative score indicating quantitative bias. Table IX.8 shows clearly that there is a consistent difference between boys and girls in the four boards, this difference being entirely consistent with the results described in the last paragraph.

Table IX.8 Mean and standard deviation of verbal/quantitative bias[a] by sex and GCE board

GCE board	MEAN BIAS		STANDARD DEVIATION OF BIAS	
	Boys	*Girls*	*Boys*	*Girls*
1	−1·735	0·292	6·442	6·001
2	−2·462	0·273	5·946	6·156
3	−2·157	0·074	6·287	6·202
4	−2·007	−0·292	5·722	5·729

[a] A positive score indicates verbal bias, a negative score indicates quantitative bias.

To find that sex differences exist in attainment as measured by GCE O level is one thing, to explain them is another. The evidence of the last paragraph suggests that at 16+ boys tend to have greater ability to tackle problems of a quantitative nature and that girls do better when the problems are verbal. So, if Test 100 is in fact an aptitude test rather than an achievement test (in so far as it is possible to make this distinction), it is reasonable to suggest that the boys in the samples tend to have superior quantitative ability and the girls superior verbal ability.

However, we are also inclined to look for explanations in the field of social psychology. Undoubtedly people achieve more when they are well motivated and it is likely that many of the candidates will have their sixth-form career in mind when sitting the examinations and that they will put their best efforts into the subjects that they consider to be useful or interesting. It is clear from the analysis of leavers who have sat for A level GCE that boys tend to have offered science subjects more often than arts subjects, and vice versa for girls (see Table 13 in *Statistics of Education 1969*[4]). Further, the importance of expectations have been cogently argued by D. A. Pidgeon in *Teacher Expectation and Pupil Performance*,[5] so it would not be unreasonable to suppose that these play a part in determining success in GCE O level – one such expectation being that girls are not expected to be good at mathematics.

Patterns of severity

Earlier in this chapter, sex grade differences were estimated, among other methods, by the difference (severity for girls only) minus (severity for boys only) for the ten subjects under consideration. There is, however, another way of looking at sex differences in the grades obtained. This is to see how the severities of the different subjects compare with each other when they are calculated (*a*) for the boys and (*b*) for the girls only. The simplest method of comparison is to plot the severities for boys only and girls only against the same axis so that the rank order and differences can easily be seen. (For a slight reservation about this method of comparison see pages 39 and 74.) Figure 6 shows this done for board 2 (using the data in Table C.9, p. 110) and is reasonably representative of the four boards which have been studied. Data in Tables C.8, 10 and 11 (pp. 109–111) will enable the reader to draw similar figures for the other boards.

A simpler, but cruder, method of making the comparison is to tabulate the rank order of severity for boys and girls. This has been done for boards 1 to 4 and the results are presented in Table IX.9.

As may be seen from Table IX.9:

a In every board, English language, English literature and French appear higher in the rank orders of severity for girls than for boys;

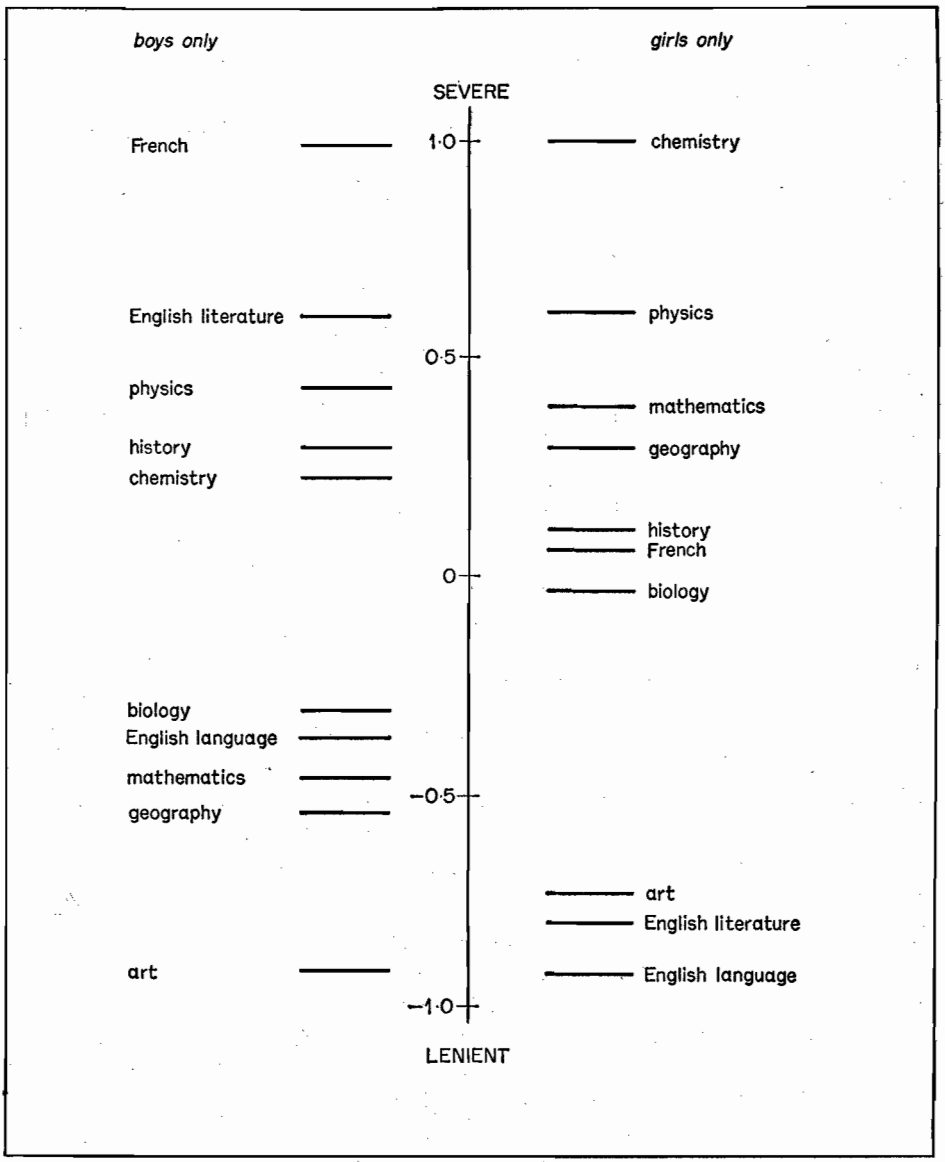

Fig. 6 *Sample estimates of severity for boys and girls separately in GCE board 2*

b With the numbers of exceptions in parentheses, geography (1), physics, mathematics and chemistry (1) appear lower in the rank orders of severity for girls than for boys.

Reference to Table IX.6 (and Tables C.5–7, pp. 108–9) will show that most of this information is contained under the column for method **ii**. What cannot be seen from that table is the rank order of the subjects for the two sexes in the

Table IX.9 Rank orders of severity by sex and GCE board[a]

	GCE BOARD							
	1		2		3		4	
Subject	*Boys*	*Girls*	*Boys*	*Girls*	*Boys*	*Girls*	*Boys*	*Girls*
Art	10	8	10	8	10	8	9	9
Biology	2	3	6	7	5	6	5	6
Chemistry	1	1	5	1	4	1	2	1
English language	7	10	7	10	7	10	8	10
English literature	4	9	2	9	2	9	3	7
French	3	6	1	6	1	4	1	3
Geography	9	5	9	4	8	7	7	8
History	6	7	4	5	3	3	4	4
Mathematics	8	4	8	3	9	5	10	5
Physics	5	2	3	2	6	2	6	2

[a] Ranked from 1 (most severe) to 10 (least severe)

four boards. There is quite good agreement between the rank orders of the boys in the four boards (mean correlation 0·83); that for the girls is also fairly good (0·83). But if we compare the rank order of the two sexes in each of the four boards (mean correlation 0·35), we find that this is much lower.

Other methods and data

Tables C.8–11 (pp. 109–111) give the reader an opportunity to compare the four main methods presented in this report when used for each sex within the four GCE boards. It will be seen that the agreement between the methods for each sex separately is just as good as that for boys and girls combined. There is no justification for supposing that the conclusions reached in Chapter VI would not apply to the methods if they were used with each sex separately.

Table C.12 (p. 111) provides estimates of severity by sex for the CSE board data forming the basis for Chapter VII. It may be seen that these data support strongly the evidence presented above for those subjects which are common to both sectors.

Severity and the calibre of candidates

In Chapter VIII the relation between the severity of a subject and the calibre of the candidates attempting it was investigated by means of a graph (Figure 5, p. 67). Since different values of severity have been found for boys and girls, it is desirable to check to see whether a similar relation holds when the analysis is carried out separately for each sex. In Figures 7 and 8 severity has been plotted against mean ability for boys and girls in board 2, using the same measure of calibre as in Figure 5. It will be seen that again there is a tendency for high values of severity to be associated with high abilities and low values of severity with low abilities. (The reader is reminded that the ANOVA measure of calibre is based on the GCE grades, and so low values correspond to high ability.)

Similar tendencies are present in the other three GCE boards and the reader will be able to plot graphs from the data in Tables C.13 and 14 (p. 112).

Severity in respect of both sexes

It must be agreed that there is considerable evidence in favour of the existence of sex differences in the grades obtained in the samples of candidates studied. Since the values of the severity obtained for the two sexes are clearly not even approximately equal, it is only reasonable to question whether it is appropriate to calculate severities for the sample as a whole, ignoring the sex of a candidate. It was suggested earlier that examiners are not concerned whether a particular candidate is a girl or a boy and so, from the point of view of a board, the concept of an overall severity is much more useful than severities applicable only to the two sexes separately. Accordingly, an attempt was made to approximate the overall severity in a board from a knowledge of the separate severities of the boys and girls in that board.

It is clear that the ratio of the number of boys to the number of girls offering a particular subject is relevant since, if no candidates of one sex took the subject, the overall severity would be that of the other sex. Accordingly, the weighted mean of the severities of the two sexes was calculated. To do this, the boys' severity was multiplied by the number of boys offering the subject, added to the same calculation for the girls and divided by the total number of candidates (boys plus girls) in the subject. The weighted means, together with the overall severities are shown in Table IX.10 for boards 1 to 4.

There is good agreement between the overall severity and its estimates for board 1, but the figures are less good as one looks across from left to right. (The mean differences for boards 1 to 4 are respectively 0·2, 0·4, 0·6 and 0·9.) In general the estimates are less than the observed overall severities, but in physics

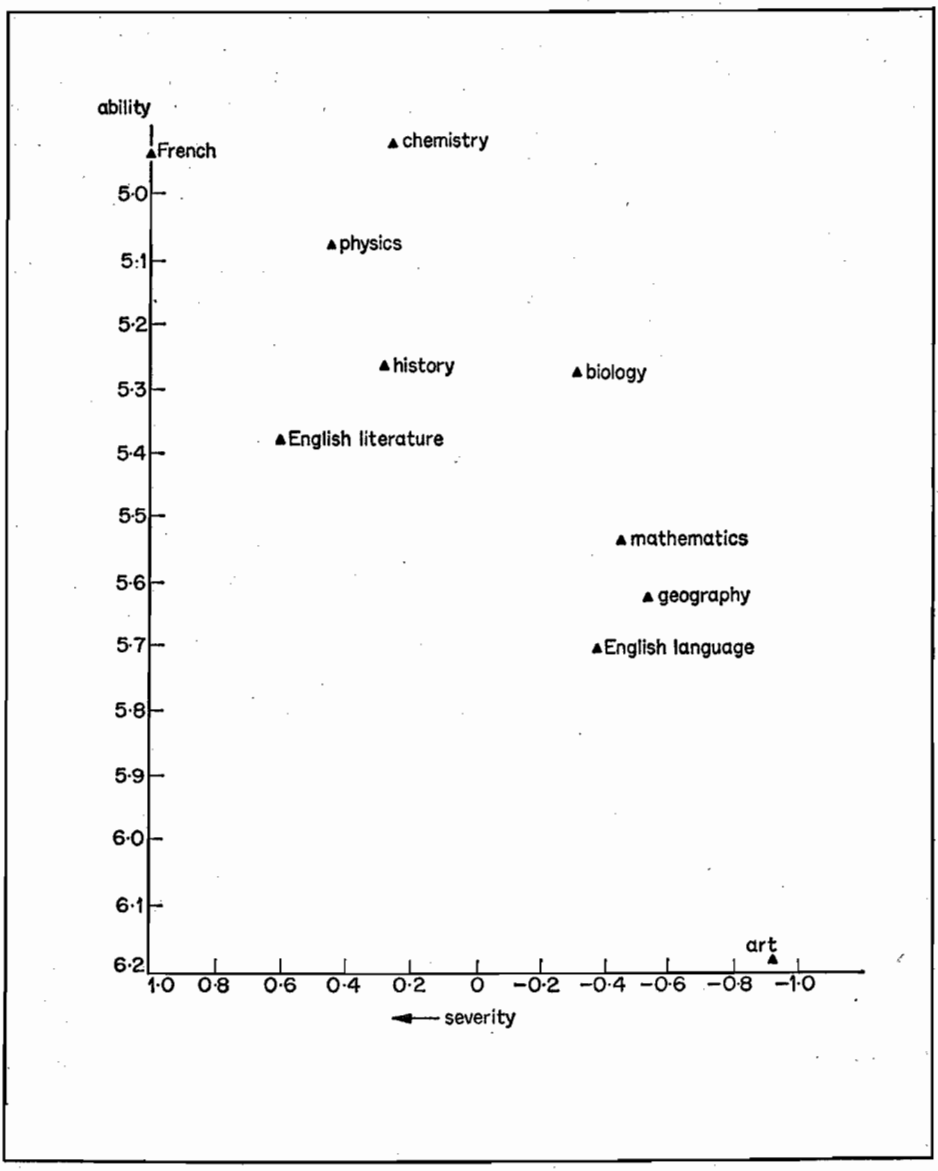

Fig. 7 *Plot of ability against severity for boys in GCE board 2* (ANOVA *method*)

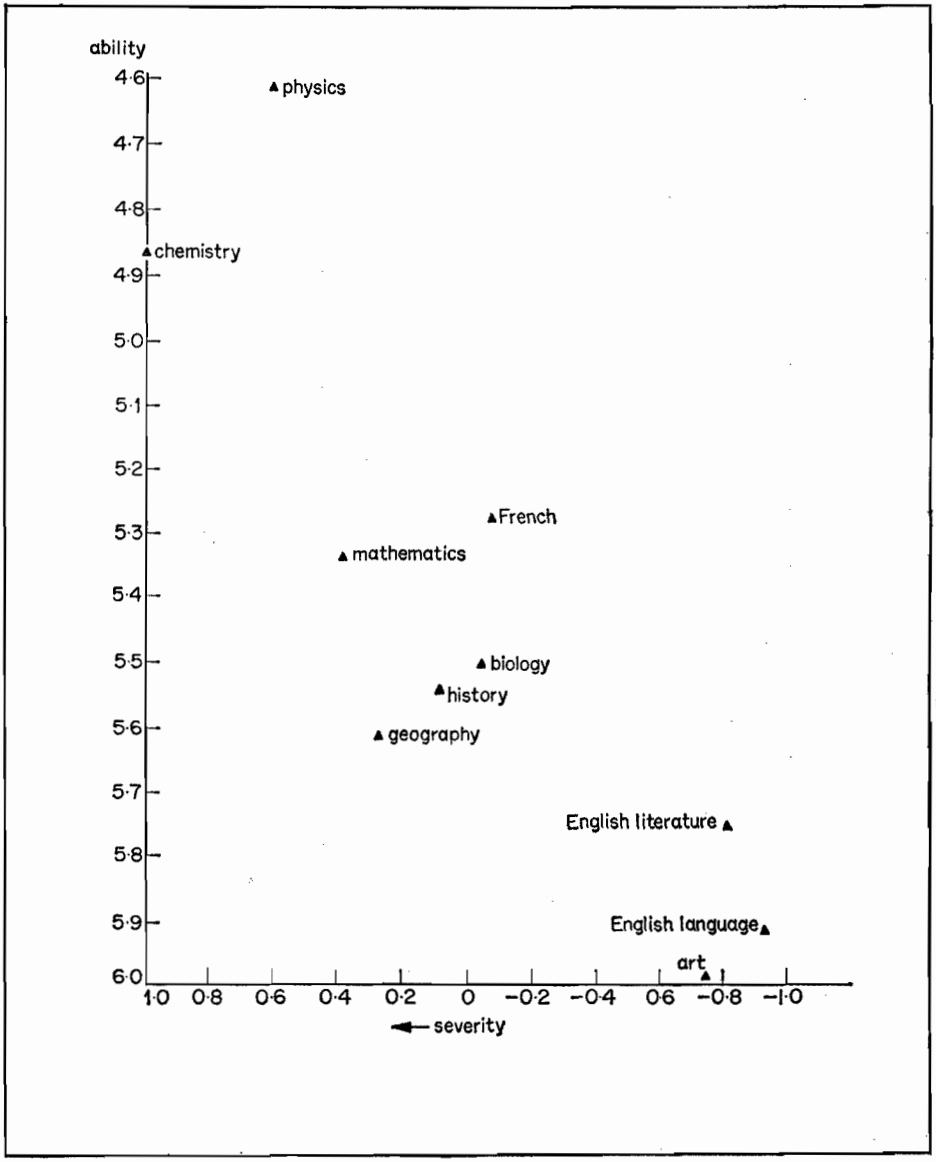

Fig. 8 *Plot of ability against severity for girls in GCE board 2* (ANOVA *method*).

Table IX.10 Overall severities and estimates by board[a]

	GCE BOARD							
	1		2		3		4	
Subject	*O*	*E*	*O*	*E*	*O*	*E*	*O*	*E*
Art	−0·68	−0·78	−0·69	−0·78	−0·63	−0·74	−0·78	−0·88
Biology	0·58	0·56	−0·04	−0·10	0·23	0·18	0·10	0·01
Chemistry	0·83	0·86	0·53	0·54	0·71	0·72	1·03	0·98
English language	−0·95	−0·96	−0·64	−0·70	−0·67	−0·74	−0·74	−0·86
English literature	−0·28	−0·29	−0·29	−0·34	−0·04	−0·10	0·01	−0·08
French	0·27	0·26	0·43	0·40	0·63	0·56	0·47	0·38
Geography	−0·20	−0·24	−0·01	−0·05	−0·54	−0·61	−0·25	−0·33
History	0·17	0·15	0·21	0·16	0·50	0·43	0·18	0·09
Mathematics	−0·09	−0·12	0·04	0·01	−0·40	−0·46	−0·47	−0·57
Physics	0·36	0·38	0·46	0·48	0·22	0·21	0·46	0·40

[a] O: overall severity calculated by the method of Chapter V.
E: estimate weighted mean of the severities for the two sexes.

and chemistry this trend is reversed or the difference is less than the average for the board concerned. These are subjects with a much larger than average proportion of boys and in which the overall severity is generally high, but the reason for the trend – and its occasional reversal – is not at all clear.

It is suggested that, in the present state of our knowledge, the correspondence between the observed overall severities and their estimates is sufficiently close for use to be made of the former. It is apparent, however, that the ratio of the number of boys and girls in the sample will make a difference to the values obtained for the overall severities. For the present, it appears that it would be helpful to think of them as the weighted means for the two sexes. A consequence of this conclusion is that it is essential to obtain a sample representative of the proportions of the two sexes when any sampling procedure is being undertaken.

Sufficient evidence is available to indicate considerable sex differences from the standpoint of subject comparability. In some subjects such as English language, English literature and French the differences are surprisingly large. There is also some basis for stating that the pattern of differences between the sexes is not limited to the GCE sector only but, as in the case of the overall pattern of subject comparability (Chapter VIII), is common to both the GCE sector and the CSE sector.

Such evidence is difficult to ignore – especially when the results across boards are so consistent – yet there seems to be little awareness of the presence of such

sex differences in the field of examining at 16+. If the awareness is there, it is not made explicit in any form, although Forrest noted the consistent 'over-performance' of girls in comparison with boys.[2] This and other topics are considered in the following chapter in order to draw together the findings of this report.

References

1. D. L. NUTTALL, *The 1968 CSE Monitoring Experiment* (Schools Council Working Paper 34). Evans/Methuen Educational, 1971.
2. G. M. FORREST, *Standards in Subjects at the Ordinary Level of the GCE, June 1970* (Occasional Publication 33). Joint Matriculation Board, Manchester, 1971.
3. F. T. TYLER, 'Individual and sex differences' in *Encyclopedia of Educational Research*, ed. C. W. Harris. Macmillan, New York, for the American Educational Research Association, 3rd edn, 1960.
4. DES, *Statistics of Education, 1969*, Vol. 2: School Leavers, GCE and CSE. HMSO, 1971.
5. D. A. PIDGEON, *Teacher Expectation and Pupil Performance*. NFER, Slough, 1970.

X. Discussion and conclusions

This chapter discusses a number of the issues raised earlier, before presenting the conclusions derived from the results of the research. One of the main issues is that of a consensus standard, since ostensibly the position of the consensus standard determines the nature of any adjustments that should be made to the standards of an individual subject to render it comparable (if it is accepted that the results reveal a genuine lack of comparability of standards between subjects). A second related issue is that of sex differences, since these may also affect the nature of any adjustments that might be made. The third issue discussed here concerns the validity of the assumptions that are made by the various methods put forward for investigating subject comparability.

The consensus standard and adjustment

In the first investigation of comparability of standards between CSE examining boards[1] an attempt was made to use the definition of CSE grade 4 to predict the expected mean grade for a given mean test score of a sample of candidates from any board. In other words, an external criterion was used to determine the 'correct' standard. In later studies (see, for example, *The 1968 CSE Monitoring Experiment*[2]) the 'correct' standard was defined to be the consensus standard of the fourteen CSE boards, that is, the average of the mean grades of each board in any given subject. Each board was given equal weight: no account was taken of the fact that the entry in English, for example, in some boards is more than double that of other boards. Estimates of severity were derived with respect to that consensus and suggested adjustments were such that the standards of a deviant board would be brought into alignment with the consensus of all boards.

The consensus standard used by the regression, guideline and subject-pair methods in this report are of the same type as employed in board comparability research. No greater weight is given to the mean grade in English (language) even though it is more popular than other subjects. The consensus is thus an average of subject mean grades. A consequence of this approach is that the estimates of severity always sum to zero. In contrast, the consensus used by the UBMT method is a weighted consensus (an average of candidates' mean grades) and the severity estimates do not sum to zero, although they may be adjusted to do so. The definition of the consensus standard in the ANOVA method is not as

clear: the fact that the estimates sum to zero is not a result of the definition of the consensus, but a mathematical convenience.

The definition of a consensus standard does not, however, present a problem because any consensus standard cannot be considered the 'correct' standard in the same way that it may be in the study of board comparability. In subject comparability, it is not the relation of one subject to the consensus that is important but the relation of one subject to each of the other subjects. Any subject could be taken as the baseline and adjustments would be made in relation to the standards of that subject. For example, English language might be taken as the baseline; the estimates of severity could be adjusted so that the standards of English language were held to be the 'correct' standard. The effect of this adjustment is shown for GCE board 2 in Table X.1.

Table X.1 Estimates of mean grade severity for GCE board 2 with respect to standards in English language (ANOVA method)

Subject	*Estimate of severity*
Art	−0·05
Biology	0·60
Chemistry	1·17
English language	—
English literature	0·35
French	1·07
Geography	0·63
History	0·85
Mathematics	0·68
Physics	1·10

This discussion may be summarized by the statement that the zero point of the severity scale is purely arbitrary. It follows that the decision as to which point is used as the baseline for any adjustment rests on considerations completely external to the methods employed.

The decision to adjust standards in different subjects to achieve subject comparability has ramifications for other aspects of comparability. For example, if GCE board 2 were to make the adjustments shown in Table X.1 in its 1974 examinations, comparability of standards between years would be disturbed in every subject except the baseline subject, English language. In French, physics and chemistry the mean grade would improve by more than one grade and the pass rate might be expected to improve by at least 10 per cent. It would therefore

become considerably easier to obtain a pass in these subjects in 1974 than it was in previous years.

Unilateral action by one board in making such adjustments would also disturb comparability of standards between boards and between sectors. If GCE board 2 made the adjustments given in Table X.1, its standards in French, physics and chemistry in particular would be lenient with respect to the standards of the other GCE boards and it would be easier to obtain an O-level pass in these subjects than a CSE grade 1, on the assumption that standards within the GCE sector and across the sectors are currently equivalent.

The achievement of equivalent standards in all subjects is therefore a matter which would require concerted action by all examining boards at 16+. There would appear to be no way of both securing equivalence of subject standards and maintaining comparability between years in a given subject at the time of adjustment. (Once new standards were established there would be no new problems in maintaining them.) The change in standards would be less marked overall if the baseline were taken as the consensus of all subjects rather than as an extreme subject such as English language.

Comparability between the sexes

A further complication concerning adjustments arises because of the sex differences in attainment discussed in Chapter IX. If adjustments were made to standards based on, for example, the deviations shown in Table X.1 which were calculated on grades for both sexes combined, a re-analysis of the adjusted data would show no differences between subjects for both sexes combined. If, however, the adjusted data were re-analysed for each sex separately, differences between subjects would still be apparent since the pattern of differences between subjects is not the same for the two sexes (see Figure 6, p. 79). Thus, the probabilities of a large group of boys obtaining the same pass rate in two subjects would not be equal, and likewise for a large group of girls, although for both groups together the probabilities would be equal.

The only way of overcoming this problem would be to grade boys and girls separately and to achieve comparability of standards between subjects within each sex. The norms for many psychological tests are given separately for each sex, and a standardized score of 115 for a boy is not considered to mean the same as a score of 115 for a girl in terms of the abilities that each is deemed to possess. An analogous situation would arise if boys' and girls' results were graded separately. In situations where boys and girls are not competing for the same jobs separate norms are highly desirable, but where they are, separate norms create problems.

The situation in which boys and girls are competing is obviously fairly com-

mon where examination results are concerned (for example, in selection for higher education) and, since no attention is paid to the sex of a candidate when his or her script is being graded, identical treatment of the two sexes is desirable and is achieved in practice. A dilemma thus remains: one can either have identical treatment of the sexes, as at present, or equivalence between subject standards for each sex separately, but one cannot have both. In either case problems remain and the decision about which is the better would depend upon considerations of the relative frequencies with which boys and girls compete for the same place and with which candidates of the same sex compete for the same place, and more broadly of the use to which examination results are put.

The assumptions of the methods

The fact that all the methods discussed in this report make a number of assumptions has been stressed throughout and most of the assumptions have been discussed in some detail. We do not claim, for example, that the assumption that the motivation of candidates is the same in each subject is wholly valid, but we presume to doubt that the observed differences between subjects are explainable solely in terms of differences in the motivation of candidates to perform well in different subjects. The same is true of most of the other minor assumptions, which are therefore not discussed further here, but two key assumptions require consideration.

The first, discussed at some length in Chapters I and VIII, is that the grading system is norm-referenced rather than criterion-referenced. If the grading system is criterion-referenced and is accepted as such, then the issue of subject comparability is irrelevant. If it is norm-referenced or is perceived as such by users, subject comparability becomes an issue of importance to users of examination results. We argue that the system is essentially norm-referenced and certainly perceived as such by users (see p. 68). Norm-referencing does not imply that grading needs to be done or is done statistically; all the important grade boundaries are in fact decided on the basis of the quality of work presented. But we argue that these criteria for grade boundaries have evolved from essentially statistical (i.e. norm-referenced) definitions, the definition of CSE grade 4 being the most obvious.

If it were the case that the grading system were entirely norm-referenced, some of the minor assumptions would become unnecessary. For example, attainment in a subject is a function both of ability and motivation: the fact that pupils studying one subject were more highly motivated than those studying another subject would raise the average level of attainment in the first subject above that of the second, but that average level would still be certified as CSE grade 4 in both cases.

The second key assumption is that the shapes of the distributions of grades is the same in each subject. It was argued in Chapter III that the evidence in the GCE sector supports this assumption, and reference to the annual reports of CSE boards shows that the distributions of grades in large-entry subjects are very similar; it is the mean grades that differ, not the general shapes of the distributions. The apparent validity of this assumption lends support to the validity of the first assumption: if the grading system were criterion-referenced there would be no reason to expect such marked similarity in the grade distributions.

It has been suggested that it is also necessary to assume that the underlying distribution of attainment in the population of 16-year-olds has the same shape and characteristics in every subject, an assumption which would be impossible to validate. But this assumption is unnecessary if the validity of the two key assumptions are accepted. In effect, in an essentially norm-referenced system the measured attainment of candidates is forced into a largely predetermined distribution. The only assumptions that are therefore necessary concern the nature of the grading system, rather than any fundamental assumptions about the nature of attainment in different subjects.

Conclusions

If the validity of the assumptions is accepted, the evidence presented in this report leads to the conclusion that there are differences between the mean grades awarded in different subjects which cannot be explained in terms of differences in the calibre of candidates entering for these subjects and which are therefore due to differences in the grading standards employed in different subjects. Since the data relate in the main to the examinations of 1968 and are based on samples whose representativeness is unknown, the magnitude of the differences in standards should not be taken to be descriptive of the current situation. The direction of the difference is, however, sufficiently consistent within sectors, across sectors and between years to suggest that currently, as in 1968, English language and literature and possibly art are likely to be leniently graded with respect to most other subjects, and that French, the physical sciences and mathematics (at least in the CSE sector) are severely graded compared with most other subjects.

There is one case where the conclusions may be stated much more strongly: the case of the results from the population of candidates in one CSE board in 1971. The data are more recent and problems of sampling candidates (and consequent sampling error) are irrelevant, i.e. the differences between subjects are real differences. When the values of severity are compared for English and arithmetic (the two most deviant subjects), English appears 0·94 grades more

lenient than arithmetic by the subject-pair method and 1·07 grades more lenient by the UBMT method. A difference of this magnitude means that if the standards in arithmetic were brought into line with those of English, every arithmetic candidate except those actually awarded grade 1 (i.e. some 2300 candidates) would have been awarded a grade which would be one better than the grade they actually received. If the standards of all subjects were adjusted to that of English in the same way, it follows that several thousand candidates would have achieved better grades in one CSE board alone.† This indicates the extent of the lack of comparability between subjects in this board in 1971; the results of the research reported here suggest a similar position in 1968 in the other boards studied, both CSE and GCE.

More importantly in the context of this report, the three methods employing no external reference instrument demonstrate a high degree of consistency in the results they provide. As argued in Chapter VI, the choice of method will consequently depend upon the size of the samples used and the computing facilities available rather than on theoretical grounds. All three are to be preferred to a method employing an external reference instrument, both on theoretical grounds and on the grounds of economy. Nevertheless an external reference instrument is essential in any investigation that goes beyond the boundaries of any one board and a scholastic aptitude test has been shown to be of value in such investigations.[2,3] Some way of linking internal and external measures of calibre would allow the simultaneous study of many aspects of comparability.

It has become apparent to us in writing this report that the comparability of standards between subjects is a most problematical issue, both from the theoretical point of view and from the point of view of its implications for current examining practice. We therefore end not with a set of tidy conclusions, but with a number of questions which are posed by the report and which we hope will stimulate discussion.

Are the existing grading systems at 16+ essentially norm-referenced or criterion-referenced?

Do the observed differences between subjects reflect differences in grading standards?

If so, are the differences in grading standards large enough to warrant adjustment of standards, at least in some subjects?

To which baseline should these adjustments be made?

Which is preferable: identical treatment of the sexes as regards examina-

† If the consensus standard of all subjects were used as a baseline, *a similar number* of grade changes would occur, but half would be upwards and half downwards.

tions, as at present, or equivalence between standards in different subjects for each sex separately (remembering that one cannot have both)?

References

1. Schools Council, *The 1965 CSE Monitoring Experiment* (Working Paper No. 6, Parts I and II). HMSO, 1966.
2. D. L. NUTTALL, *The 1968 CSE Monitoring Experiment* (Schools Council Working Paper 34). Evans/Methuen Educational, 1971.
3. LARRY S. SKURNIK, *Monitoring Grade Standards in English* (Schools Council Working Paper 49). Evans/Methuen Educational, 1974.

Appendices

Appendix A Further details of the methods

The regression method

For each candidate in the sample, the raw data comprise his total score on Test 100 and the grades achieved in as many of the ten subjects as he attempted. The sub-sample of candidates attempting each subject in turn is isolated and their mean test score and mean grade are calculated. The data are also analysed by an analysis of variance and co-variance (between and within schools) to produce, inter alia, the pooled within-school regression coefficients and the pooled within-school correlation coefficient between test scores and grades.

The mean pooled within-school regression coefficient of grade on test score ($\bar{b}$) is then calculated (strictly speaking, this should only be done if there are no significant differences between the regression coefficients). The means of the mean test scores by subject and of the subject mean grades ($\bar{y}$ and $\bar{u}$ respectively) are also calculated. In the calculation of these grand means, it should be noted that no account is taken of the differing numbers of candidates in each sub-sample, i.e. the consensus standard is unweighted. The overall regression line of grade-on-test, with slope $\bar{b}$ and passing through the point $\bar{u}$, $\bar{y}$, therefore has the equation:

$$u = \bar{u} + \bar{b}(y - \bar{y}) \tag{1}$$

For subject j we have an observed mean grade of u_j and an observed mean test score of y_j. The predicted mean grade u_j for a given mean test score of y_j is, from equation (1),

$$u_j' = \bar{u} + \bar{b}(y_j - \bar{y}) \tag{2}$$

and the estimate of severity is the difference between the observed and predicted mean grades, i.e. $u_j - u_j'$.

The guideline method

As noted in Chapter II, the regression method has been criticized on technical grounds by Please.† Peaker‡ accepts these criticisms, noting that in the

† 'The 1965 CSE Monitoring Experiment: a comment', *Educational Research*, **13** (June 1971), 233–5.

‡ 'The 1965 CSE Monitoring Experiment: a reply [to comment by Please]', *Educational Research*, **13** (June 1971), 235–6.

regression method as outlined above all the error is assumed to lie in the grades (since the grade-on-test regression line is used; the reverse would be true in the case of the test-on-grade regression line). In structural regression, of which the guideline is a special case, the error is more reasonably assumed to lie partly in the grades and partly in the test scores. The structural regression line therefore lies somewhere between the two regression lines, but the problem is to determine the proportion in which to divide the error to fix the slope of the line.

In the absence of any evidence, the most appropriate course of action would seem to be to split the error evenly, giving the structural regression line which bisects the angle between the two regression lines if the two variables are standardized. Peaker argues, however, that in the case of examinations at 16+ indirect evidence concerning the appropriate split is available. This arises because information is available about two distinct clusters of candidates, those entered for GCE O level and those entered for CSE. The most suitable line, here called the guideline, is the line through the two points corresponding to the general means of the GCE and CSE sectors.

Along the vertical axis, the distance between the two points in terms of test scores presents no difficulty since the same test is used in both sectors. Along the horizontal axis, the distance in terms of grades presents a problem since the grade scales in the two sectors are different. In the studies reported here, it is assumed that the boundary between grades 1 and 2 in CSE corresponds to the boundary between GCE grades 6 and 7 and further that a GCE grade is 0·6 the width of a CSE grade. (Peaker presents evidence supporting these two assumptions and unpublished work indicates that varying the second assumption slightly makes very little difference to the slope of the guideline.)

The slope of the guideline may now be calculated as follows: let $\bar{u}_C$ and $\bar{y}_C$ be the grand mean CSE grade and Test 100 score, and $\bar{u}_G$ and $\bar{y}_G$ be the grand mean GCE grade and Test 100 score. Then the vertical distance between the two grand means is simply $\bar{y}_G - \bar{y}_C$, while the horizontal distance is given by the sum of the distances between each mean and the overlap point (the CSE grade 1/2 boundary or the GCE grade 6/7 boundary) after correction of one distance into the metric of the other. Thus, in terms of CSE grades, the distance is:

$$(\bar{u}_C - 1) + 0{\cdot}6\,(6 - \bar{u}_G)$$

and, in terms of GCE grades, it is:

$$\frac{10}{6}(\bar{u}_C - 1) + (6 - \bar{u}_G).$$

The slope (in terms of grades/test score) is simply the horizontal distance

divided by the vertical distance. The values of $\bar{u}_C$ and $\bar{y}_C$ used in this report were 3·02 and 38·8 and were obtained by taking the means of the mean grades across all CSE boards and the mean test scores across all CSE boards in the following nine subjects: art, biology, chemistry, English language, French, geography, history, mathematics and physics (English literature being excluded since most CSE boards do not offer a Mode I examination in this subject). The means for each subject are given in Table 6 of *The 1968 CSE Monitoring Experiment* (Evans/Methuen Educational, 1971), except in the case of art where the data are unpublished.

The predicted grade for a given test score is given by equation (2) above, substituting the slope of the guideline for $\bar{b}$, the mean regression coefficient. However, the estimate of severity is no longer the difference between the observed grade and the predicted grade since evidence is now available about the appropriate division of error. Peaker demonstrates that the guideline corresponds to the structural regression line which would be obtained by dividing the error in the ratio of four to five in favour of the grades against the test scores, and this ratio was confirmed in the context of the present work.

In calculating the estimates of severity it is necessary to apply this ratio to the distance between the observed (u_j) and predicted (u_j') grades; the estimate of severity is thus $\frac{4}{9}(u_j - u_j')$. In the simplified explanation of the guideline method given in Chapter II, this feature was omitted and the 'predicted grade' used in the chemistry example was not in fact correct, although the estimate of severity was correct.

The relationship between the regression method and the guideline method

The regression line and the guideline both pass through the same point, the grand mean, but differ in slope. Because the guideline is equivalent to one of the family of structural regression lines, its slope will always be less than the slope of the regression line of grades on test scores. Figure A.1 provides a hypothetical example.

The point A appears on the lenient side of the regression line but on the severe side of the guideline, while the opposite is true of point D. For points such as A and D, the difference between the estimates of severity provided by the two methods can be very marked. In contrast, points B and C remain on the same side of the two lines; although they are considerably further horizontally from the guideline than they are from the regression line, the estimates of severity or leniency tend not to be very different since the guideline estimate is only $\frac{4}{9}$ of the distance of the point from the regression line.

In the study of subject comparability, a tendency has been noted (see Chapters I and II) for those subjects which attract an entry of high calibre as measured

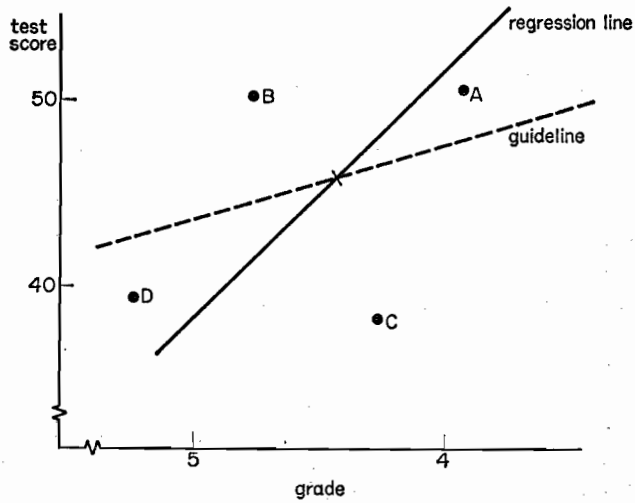

Fig. A.1 *Hypothetical example of regression line and guideline*

by the test to be those subjects which award poor grades on average, i.e. for the points to lie along the line formed by joining points B and C. This serves to explain why the observed differences between the estimates of severity provided by the two methods tend to be fairly small. In studies of comparability between examining boards, the evidence in *The 1968 CSE Monitoring Experiment* shows that the points tend to lie along the line formed by joining points A and D, as is to be expected since there was no evidence of a lack of comparability of standards. Large differences, including differences in direction, between the estimates of severity provided by the two methods are therefore more likely in the study of comparability between examining boards.

Appendix B Background data

Table B.1 Number of schools representing GCE boards

Subject	BOARD 1	2	3	4
Art	32	51	17	21
Biology	27	41	11	16
Chemistry	22	41	9	15
English language	31	53	22	23
English literature	27	46	18	16
French	23	40	14	14
Geography	28	50	21	18
History	24	39	13	19
Mathematics	31	50	19	20
Physics	26	44	14	16

Table B.2 Number of candidates representing GCE boards

Subject	BOARD 1	2	3	4
Art	562	765	174	269
Biology	837	1372	289	560
Chemistry	449	915	153	334
English language	1579	2828	591	798
English literature	1308	2548	502	612
French	928	1635	342	608
Geography	972	1948	448	634
History	740	1422	270	666
Mathematics	1091	1892	476	689
Physics	562	977	179	399

Table B.3 Mean subject grades ($\bar{U}$) for GCE boards and standard error

	BOARD							
	1		2		3		4	
Subject	$\bar{U}$	*SE*	$\bar{U}$	*SE*	$\bar{U}$	*SE*	$\bar{U}$	*SE*
Art	5·25	0·21	5·27	0·16	4·99	0·41	5·16	0·24
Biology	5·80	0·30	5·35	0·15	5·26	0·33	5·44	0·29
Chemistry	5·63	0·30	5·44	0·21	5·26	0·31	5·85	0·26
English language	4·61	0·23	5·14	0·11	5·24	0·23	5·15	0·24
English literature	5·05	0·25	5·30	0·15	5·40	0·22	5·69	0·20
French	4·97	0·20	5·56	0·16	5·85	0·28	5·72	0·29
Geography	5·03	0·18	5·57	0·15	5·02	0·19	5·26	0·23
History	5·13	0·30	5·61	0·17	5·67	0·45	5·66	0·16
Mathematics	5·08	0·18	5·45	0·14	5·00	0·20	5·22	0·17
Physics	5·60	0·27	5·44	0·17	5·19	0·35	5·52	0·25
Mean	5·21	0·24	5·41	0·16	5·29	0·29	5·46	0·23

Table B.4 Mean Test 100 scores ($\bar{Y}$) for GCE boards and standard error

	BOARD							
	1		2		3		4	
Subject	$\bar{Y}$	*SE*	$\bar{Y}$	*SE*	$\bar{Y}$	*SE*	$\bar{Y}$	*SE*
Art	45·2	1·3	47·3	0·8	46·8	1·0	48·5	1·5
Biology	50·7	1·0	50·6	0·7	53·1	1·9	53·2	1·1
Chemistry	56·1	1·0	55·0	0·6	58·0	1·1	58·4	1·0
English language	50·3	0·9	49·1	0·7	48·6	1·2	50·0	1·3
English literature	51·4	1·0	50·0	0·7	50·1	1·5	52·3	1·1
French	54·2	0·7	52·7	0·6	52·8	1·1	55·3	0·9
Geography	51·8	0·8	50·7	0·6	50·4	1·1	54·3	1·1
History	51·9	1·0	51·1	0·8	50·9	2·1	53·1	1·1
Mathematics	52·5	0·8	52·5	0·5	52·3	0·8	53·6	0·8
Physics	54·6	1·2	55·6	0·5	57·3	1·2	58·7	0·9
Mean	51·9	1·0	51·5	0·7	52·0	1·3	53·7	1·1

Table B.5 Correlations between Test 100 scores and GCE grades (pooled within-school estimates)[a]

Subject	BOARD 1	2	3	4
Art	0·168	0·245	0·112	0·076
Biology	0·390	0·440	0·464	0·429
Chemistry	0·293	0·354	0·436	0·453
English language	0·349	0·358	0·320	0·341
English literature	0·187	0·330	0·267	0·326
French	0·316	0·383	0·314	0·342
Geography	0·407	0·393	0·357	0·419
History	0·246	0·259	0·288	0·289
Mathematics	0·418	0·525	0·476	0·483
Physics	0·344	0·404	0·372	0·485
Mean	0·312	0·369	0·341	0·364

[a] Since good examination performance is represented by a low numbered grade, the signs of the correlation coefficients have had to be changed. This procedure is adopted throughout the report in presenting correlations to avoid confusion.

Table B.6 Number of candidates, mean grade ($\bar{U}$), mean Test 100 score ($\bar{Y}$), mean verbal Test 100 score ($\bar{V}$) and mean Test E2 score ($\bar{E}$) for a sub-sample from GCE board 5

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	$\bar{V}$	$\bar{E}$
Art	61	5·12	39·8	20·1	29·6
Biology	138	5·95	45·2	22·1	32·4
Chemistry	45	6·11	48·5	24·0	33·0
English language	202	4·75	44·0	21·8	31·5
English literature	184	5·34	44·6	21·9	32·2
French	100	4·96	49·2	24·3	35·4
Geography	115	5·64	44·7	22·0	31·2
History	102	5·16	43·7	21·9	32·3
Mathematics	118	4·65	48·0	23·3	33·4
Physics	42	5·52	51·7	25·3	33·9
Mean	—	5·32	45·9	22·7	32·5

Table B.7 Correlations between grade (U), Test 100 score (Y), Verbal Test 100 score (V) and Test E2 score (E) for a sub-sample from GCE board 5

Subject	*U and Y*	*U and V*	*U and E*
Art	0·358	0·328	0·313
Biology	0·535	0·484	0·380
Chemistry	0·702	0·718	0·614
English language	0·419	0·455	0·597
English literature	0·297	0·324	0·401
French	0·383	0·415	0·450
Geography	0·309	0·317	0·250
History	0·420	0·402	0·440
Mathematics	0·493	0·298	0·355
Physics	0·654	0·655	0·532
Mean	0·457	0·440	0·433

Table B.8 Number of candidates, mean grade ($\bar{U}$), mean Test 100 score ($\bar{Y}$), mean verbal Test 100 score ($\bar{V}$) and mean Test E2 score ($\bar{E}$) for a sub-sample from CSE board 15

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	$\bar{V}$	$\bar{E}$
Art	44	2·86	32·7	16·1	18·5
Biology	38	2·84	34·6	17·6	23·9
Chemistry	50	3·14	37·5	19·0	24·5
English	160	2·85	33·3	16·5	19·9
French	26	2·35	37·5	18·7	27·3
Geography	82	3·06	34·3	17·0	20·2
History	74	3·26	33·6	17·1	20·4
Mathematics	164	3·49	33·6	17·8	22·4
Physics	78	3·44	36·1	18·3	22·2
Mean	—	3·03	34·8	17·6	22·1

Table B.9 Correlations between grade (U), Test 100 score (Y), Verbal Test 100 score (V) and Test E2 (E) for a sub-sample from CSE board 15

Subject	*U and V*	*U and V*	*U and E*
Art	−0·002	0·085	0·038
Biology	0·471	0·456	0·291
Chemistry	0·457	0·474	0·373
English	0·348	0·373	0·479
French	−0·195	−0·334	0·039
Geography	0·331	0·299	0·241
History	0·459	0·395	0·224
Mathematics	0·416	0·216	0·267
Physics	0·507	0·513	0·528
Mean	0·310	0·350	0·276

Table B.10 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for boys only in GCE board 1

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlation*
Art	251	5·30	43·83	0·061
Biology	258	6·42	53·42	0·388
Chemistry	329	6·04	55·84	0·305
English language	702	5·49	51·56	0·488
English literature	537	5·94	53·65	0·285
French	343	5·55	57·07	0·258
Geography	396	4·93	53·48	0·332
History	257	5·54	54·31	0·292
Mathematics	544	5·24	52·16	0·418
Physics	442	5·86	53·81	0·454
Mean	—	5·63	52·91	0·328

Table B.11 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for girls only in GCE board 1

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlations*
Art	311	5·22	46·32	0·233
Biology	579	5·52	49·62	0·471
Chemistry	120	4·50	56·97	0·528
English language	877	3·91	49·42	0·478
English literature	771	4·44	49·94	0·329
French	585	4·63	52·52	0·408
Geography	576	5·09	50·67	0·485
History	483	4·91	50·70	0·340
Mathematics	547	4·92	52·86	0·552
Physics	120	4·65	57·57	0·401
Mean	—	4·78	51·66	0·423

Table B.12 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for boys only in GCE board 2

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlation*
Art	252	5·29	50·0	0·067
Biology	335	4·97	55·0	0·359
Chemistry	558	5·17	56·4	0·258
English language	1073	5·33	53·2	0·455
English literature	815	5·99	54·8	0·301
French	591	5·94	56·7	0·311
Geography	805	5·10	53·7	0·374
History	496	5·55	55·4	0·239
Mathematics	857	5·10	54·0	0·506
Physics	710	5·51	56·0	0·384
Mean	—	5·40	54·5	0·325

Table B.13 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for girls only in GCE board 2

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlation*
Art	513	5·27	46·2	0·187
Biology	1037	5·48	49·2	0·466
Chemistry	357	5·87	53·0	0·452
English language	1755	5·02	46·8	0·435
English literature	1733	4·97	47·9	0·376
French	1044	5·35	50·6	0·417
Geography	1143	5·91	48·6	0·440
History	926	5·64	48·9	0·268
Mathematics	1035	5·75	51·3	0·570
Physics	267	5·23	54·9	0·406
Mean	—	5·45	49·7	0·402

Table B.14 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for boys only in GCE board 3

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlation*
Art	75	4·68	46·40	0·008
Biology	91	5·08	58·13	0·514
Chemistry	86	5·09	59·51	0·443
English language	263	5·85	51·64	0·434
English literature	185	6·31	54·49	0·237
French	106	6·31	57·51	0·363
Geography	208	4·78	52·70	0·295
History	103	5·79	56·15	0·220
Mathematics	236	4·67	53·50	0·451
Physics	134	5·43	56·82	0·487
Mean	—	5·40	54·69	0·345

Table B.15 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for girls only in GCE board 3

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlation*
Art	99	5·23	47·21	0·126
Biology	198	5·35	50·83	0·399
Chemistry	67	5·48	56·06	0·597
English language	328	4·75	46·75	0·493
English literature	317	4·86	48·26	0·351
French	236	5·65	50·78	0·425
Geography	240	5·24	48·86	0·363
History	167	5·59	48·96	0·365
Mathematics	240	5·32	51·20	0·489
Physics	45	4·47	58·87	0·461
Mean	—	5·19	50·78	0·407

Table B.16 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for boys only in GCE board 4

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlation*
Art	89	5·75	49·84	0·063
Biology	177	5·58	56·93	0·311
Chemistry	218	5·66	59·76	0·377
English language	290	5·78	54·20	0·404
English literature	255	6·02	56·08	0·197
French	255	6·31	58·05	0·382
Geography	360	5·41	56·80	0·249
History	319	5·82	56·29	0·248
Mathematics	338	4·84	55·75	0·423
Physics	302	5·45	58·77	0·524
Mean	—	5·66	56·25	0·318

Table B.17 Number of candidates, mean grades ($\bar{U}$), mean Test 100 scores ($\bar{Y}$) and correlations between grade and score for girls only in GCE board 4

Subject	*No. of candidates*	$\bar{U}$	$\bar{Y}$	*Correlation*
Art	180	4·86	47·89	0·150
Biology	383	5·38	51·53	0·474
Chemistry	116	6·22	56·12	0·654
English language	508	4·79	47·71	0·507
English literature	357	5·45	49·74	0·430
French	353	5·29	53·36	0·440
Geography	274	5·07	51·04	0·543
History	347	5·51	50·31	0·335
Mathematics	351	5·59	51·68	0·505
Physics	97	5·74	58·69	0·544
Mean	—	5·39	51·81	0·460

Appendix C Further results

Table C.1 Sample estimates of mean grade severity for GCE board 1: four methods compared[a]

Subject	METHOD			
	Regression	*Subject-pair*	UBMT	ANOVA
Art	−0·47	−0·66	−0·72	−0·68
Biology	0·48	0·58	0·69	0·58
Chemistry	0·71	0·91	0·82	0·83
English language	−0·74	−1·00	−1·09	−0·95
English literature	−0·21	−0·30	−0·27	−0·28
French	0·09	0·36	0·30	0·27
Geography	−0·21	−0·29	−0·18	−0·20
History	−0·10	0·14	0·21	0·17
Mathematics	−0·10	−0·15	−0·10	−0·09
Physics	0·58	0·42	0·34	0·36

[a] Four methods only are compared since, for technical reasons, guideline estimates of severity are not available for GCE board 1.

Table C.2 Sample estimates of mean grade severity for GCE board 3: the five methods compared

Subject	METHOD				
	Regression	*Guideline*	*Subject-pair*	UBMT	ANOVA
Art	−0·74	−1·16	−0·40	−0·59	−0·63
Biology	0·05	0·17	0·12	0·23	0·23
Chemistry	0·46	1·01	0·73	0·66	0·71
English language	−0·33	−0·60	−0·77	−0·70	−0·67
English literature	−0·05	−0·26	0·08	−0·04	−0·04
French	0·62	0·45	0·59	0·69	0·63
Geography	−0·42	−0·42	−0·55	−0·61	−0·54
History	0·31	0·04	0·56	0·57	0·50
Mathematics	−0·29	−0·12	−0·52	−0·42	−0·40
Physics	0·32	0·85	0·16	0·22	0·22

Table C.3 Sample estimates of mean grade severity for GCE board 4: the five methods compared

Subject	METHOD				
	Regression	*Guideline*	*Subject-pair*	UBMT	ANOVA
Art	−0·76	−0·93	−0·49	−0·75	−0·78
Biology	−0·07	−0·09	−0·01	0·17	0·10
Chemistry	0·79	0·90	0·99	0·98	1·03
English language	−0·64	−0·72	−0·77	−0·77	−0·74
English literature	0·10	−0·08	−0·08	0·08	0·01
French	0·29	0·37	0·61	0·49	0·47
Geography	−0·15	−0·02	−0·32	−0·27	−0·25
History	0·14	0·01	0·15	0·21	0·18
Mathematics	−0·25	0·15	−0·58	−0·52	−0·47
Physics	0·48	0·75	0·50	0·37	0·46

Table C.4 Sample mean ability values[a] and estimates of severity for GCE boards 1–4 (ANOVA method)

Subject	GCE BOARD							
	1		2		3		4	
	Ability	*Severity*	*Ability*	*Severity*	*Ability*	*Severity*	*Ability*	*Severity*
Art	5·94	−0·68	5·96	−0·69	5·63	−0·63	5·93	−0·78
Biology	5·22	0·58	5·39	−0·04	5·03	0·23	5·34	0·10
Chemistry	4·80	0·83	4·92	0·53	4·55	0·71	4·83	1·03
English language	5·56	−0·95	5·78	−0·64	5·90	−0·67	5·89	−0·74
English literature	5·33	−0·28	5·59	−0·29	5·44	−0·44	5·67	0·01
French	4·70	0·27	5·13	0·43	5·23	0·63	5·25	0·47
Geography	5·23	−0·20	5·58	−0·01	5·57	−0·54	5·52	−0·25
History	4·96	0·17	5·40	0·21	5·17	0·50	5·48	0·18
Mathematics	5·17	−0·09	5·41	0·04	5·40	−0·40	5·69	−0·47
Physics	5·24	0·36	4·98	0·46	4·97	0·22	5·06	0·46

[a] $a + c_i$ (see p. 38)

Table C.5 Grade differences between the sexes by subject in GCE board 1 using four different methods

Subject	METHOD OF ESTIMATING GRADE DIFFERENCE			
	Difference between severities (ii)	*Allowance made for* c_i (iii)	*Uncontrolled difference* (i)	*Allowance made for Test 100* (iv)
Art	0·75	0·50	−0·08	−0·00
Biology	−0·35	−0·02	−0·90	−1·33
Chemistry	−0·20	0·19	−1·54	−1·43
English language	−0·80	−0·76	−1·58	−1·79
English literature	−1·03	−0·88	−1·50	−1·76
French	−0·48	−0·37	−0·92	−1·29
Geography	0·98	1·04	0·16	−0·12
History	−0·19	0·08	−0·63	−0·93
Mathematics	0·99	0·94	−0·32	−0·23
Physics	0·31	0·68	−1·21	−0·78

Table C.6 Grade differences between the sexes by subject in GCE board 3 using four different methods

Subject	METHOD OF ESTIMATING GRADE DIFFERENCE			
	Difference between severities (ii)	*Allowance made for* c_i (iii)	*Uncontrolled difference* (i)	*Allowance made for Test 100* (iv)
Art	0·82	0·73	0·55	0·57
Biology	−0·19	−0·04	0·27	−0·54
Chemistry	0·79	0·95	0·39	−0·09
English language	−1·02	−0·76	−1·10	−1·57
English literature	−1·73	−1·41	−1·45	−1·86
French	−0·94	−0·71	−0·66	−1·40
Geography	0·74	0·80	0·46	0·17
History	−0·09	0·15	−0·20	−0·77
Mathematics	1·20	1·17	0·65	0·39
Physics	0·42	0·54	−0·96	−0·71

Table C.7 Grade differences between the sexes by subject in GCE board 4 using four different methods

Subject	METHOD OF ESTIMATING GRADE DIFFERENCE			
	Difference between severities (ii)	*Allowance made for* c_i (iii)	*Uncontrolled difference* (i)	*Allowance made for Test 100* (iv)
Art	−0·65	−0·51	−0·89	−0·92
Biology	−0·18	−0·07	−0·20	−0·80
Chemistry	0·98	0·88	0·56	0·10
English language	−1·00	−0·76	−0·99	−1·60
English literature	−0·72	−0·54	−0·57	−1·04
French	−0·57	−0·45	−1·02	−1·51
Geography	−0·26	−0·15	−0·34	−0·83
History	−0·17	−0·07	−0·31	−0·68
Mathematics	1·17	1·12	0·75	0·29
Physics	1·43	1·31	0·29	0·28

Table C.8 Sample estimates of mean grade severity for GCE board 1 by sex and method

Subject	BOYS				GIRLS			
	Re-gression	*Subject-pair*	UBMT	ANOVA	*Re-gression*	*Subject-pair*	UBMT	ANOVA
Art	−1·02	−0·90	−1·26	−1·20	−0·11	−0·51	−0·38	−0·45
Biology	0·83	0·96	0·88	0·80	0·53	0·41	0·56	0·45
Chemistry	0·63	0·93	0·96	0·91	0·27	0·80	0·59	0·71
English language	−0·24	−0·65	−0·54	−0·52	−0·10	−1·25	−1·54	−1·32
English literature	0·36	0·30	0·36	0·32	−0·52	−0·78	−0·74	−0·71
French	0·23	0·62	0·53	0·56	−0·06	0·19	0·13	0·08
Geography	−0·65	−0·92	−0·88	−0·82	0·21	0·12	0·28	0·16
History	0·01	0·18	0·24	0·27	0·03	0·04	0·15	0·08
Mathematics	−0·45	−0·70	−0·65	−0·62	0·27	0·27	0·43	0·37
Physics	0·30	0·30	0·36	0·31	0·48	0·71	0·50	0·62

Table C.9 Sample estimates of mean grade severity for GCE board 2 by sex and method

Subject	BOYS				GIRLS			
	Re-gression	*Subject-pair*	UBMT	ANOVA	*Re-gression*	*Subject-pair*	UBMT	ANOVA
Art	−0·46	−0·77	−0·92	−0·90	−0·51	−0·65	−0·68	−0·72
Biology	−0·38	−0·44	−0·31	−0·31	−0·02	−0·16	0·03	−0·03
Chemistry	−0·08	0·28	0·20	0·24	0·73	1·10	0·95	1·00
English language	−0·17	−0·47	−0·40	−0·37	−0·71	−0·91	−0·95	−0·90
English literature	0·62	0·66	0·68	0·61	−0·66	−0·91	−0·83	−0·79
French	0·72	1·14	1·04	1·00	−0·02	0·16	0·05	0·06
Geography	−0·36	−0·60	−0·57	−0·53	0·36	0·24	0·38	0·29
History	0·23	0·28	0·30	0·28	0·11	0·17	0·15	0·10
Mathematics	−0·34	−0·52	−0·48	−0·45	0·45	0·38	0·44	0·40
Physics	0·23	0·44	0·46	0·43	0·27	0·59	0·47	0·61

Table C.10 Sample estimates of mean grade severity for GCE board 3 by sex and method

Subject	BOYS				GIRLS			
	Re-gression	*Subject-pair*	UBMT	ANOVA	*Re-gression*	*Subject-pair*	UBMT	ANOVA
Art	−1·48	−1·01	−1·11	−1·21	−0·18	−0·06	−0·32	−0·39
Biology	−0·01	0·37	0·18	0·31	−0·25	−0·05	0·16	0·12
Chemistry	0·13	0·24	0·25	0·37	0·84	1·27	1·13	1·16
English language	0·17	−0·34	−0·03	−0·17	−0·71	−1·21	−1·31	−1·19
English literature	0·90	1·13	1·14	0·99	−0·46	−0·67	−0·82	−0·74
French	1·17	1·25	1·21	1·21	0·55	0·25	0·36	0·27
Geography	−0·80	−1·09	−1·11	−1·01	−0·03	−0·16	−0·29	−0·27
History	0·52	0·60	0·48	0·49	0·33	0·54	0·53	0·40
Mathematics	−0·84	−1·18	−1·17	−1·07	0·26	−0·06	0·19	0·13
Physics	0·23	0·03	0·17	0·10	−0·84	0·24	0·37	0·52

Table C.11 Sample estimates of mean grade severity for GCE board 4 by sex and method

Subject	BOYS				GIRLS			
	Re-gression	*Subject-pair*	UBMT	ANOVA	*Re-gression*	*Subject-pair*	UBMT	ANOVA
Art	−0·39	−0·02	−0·47	−0·45	−0·95	−0·76	−1·04	−1·10
Biology	−0·03	0·10	0·16	0·13	−0·04	−0·19	0·02	−0·05
Chemistry	0·26	0·53	0·63	0·64	1·29	1·67	1·48	1·62
English language	−0·04	−0·35	−0·16	−0·22	−1·04	−1·21	−1·29	−1·22
English literature	0·35	0·25	0·37	0·34	−0·16	−0·55	−0·27	−0·38
French	0·79	0·82	0·75	0·71	0·07	0·29	0·14	0·14
Geography	−0·21	−0·26	−0·25	−0·22	−0·40	−0·65	−0·44	−0·48
History	0·17	0·14	0·21	0·18	−0·04	−0·08	0·07	0·01
Mathematics	−0·86	−1·31	−1·29	−1·17	0·19	−0·10	0·06	0·00
Physics	−0·02	0·10	0·04	0·05	1·10	1·50	1·28	1·48

Table C.12 Sample estimates of mean grade severity for the twenty subjects by sex and overall for the CSE board (ANOVA method)

Subject	*Boys*	*Girls*	*Overall*
Arithmetic	0·12	0·45	0·44
Art and crafts	−0·20	−0·32	−0·20
Biology	−0·23	−0·08	−0·07
Business studies	−0·55	−0·13	−0·07
Chemistry	0·28	0·79	0·44
Commerce	−0·04	0·17	−0·07
English	−0·30	−0·85	−0·48
French	0·45	−0·06	0·15
Geography	−0·47	−0·20	−0·30
German	1·27	0·21	0·56
History	−0·32	−0·32	−0·26
Home economics (Cookery and hostess)	0·07	−0·40	−0·30
Integrated science	0·01	0·38	0·13
Mathematics	−0·14	0·32	0·10
Metalwork	−0·14	—	−0·08
Needlework (Fashion)	—	−0·10	−0·04
Physics	0·01	0·53	0·10
Religious knowledge	0·68	0·07	0·32
Technical drawing	−0·07	−0·12	−0·02
Woodwork	−0·41	—	−0·34
No. of candidates[a]	829	797	1626

[a] Only candidates offering two or more subjects are included in the above analysis.

Table C.13 Sample estimates of mean grade severity by sex for GCE boards 1–4 (ANOVA method)

	GCE BOARD							
	1		2		3		4	
Subject	*Boys*	*Girls*	*Boys*	*Girls*	*Boys*	*Girls*	*Boys*	*Girls*
Art	−1·20	−0·45	−0·90	−0·72	−1·21	−0·39	−0·45	−1·10
Biology	0·80	0·45	−0·31	−0·03	0·31	0·12	0·13	−0·05
Chemistry	0·91	0·71	0·24	1·00	0·37	1·16	0·64	1·62
English language	−0·52	−1·32	−0·37	−0·90	−0·17	−1·19	−0·22	−1·22
English literature	0·32	−0·71	0·61	−0·79	0·99	−0·74	0·34	−0·38
French	0·56	0·08	1·00	0·06	1·21	0·27	0·71	0·14
Geography	−0·82	0·16	−0·53	0·29	−0·01	−0·27	−0·22	−0·48
History	0·27	0·08	0·28	0·10	0·49	0·40	0·18	0·01
Mathematics	−0·62	0·37	−0·45	0·40	−1·07	0·13	−1·17	0·00
Physics	0·31	0·62	0·43	0·61	0·10	0·52	0·05	1·48

Table C.14 Sample mean ability values[a] by sex for GCE boards 1–4 (ANOVA method)

	GCE BOARD							
	1		2		3		4	
Subject	*Boys*	*Girls*	*Boys*	*Girls*	*Boys*	*Girls*	*Boys*	*Girls*
Art	6·50	5·66	6·18	5·99	5·89	5·62	6·20	5·96
Biology	5·62	5·07	5·28	5·51	4·77	5·23	5·44	5·43
Chemistry	5·13	3·79	4·93	4·88	4·72	4·32	5·01	4·60
English language	6·01	5·23	5·71	5·92	6·02	5·94	6·00	6·01
English literature	5·62	5·15	5·38	5·76	5·32	5·60	5·68	5·83
French	4·99	4·54	4·94	5·29	5·10	5·38	5·60	5·15
Geography	5·75	4·93	5·63	5·62	5·79	5·51	5·63	5·55
History	5·27	4·83	5·27	5·55	5·29	5·19	5·64	5·50
Mathematics	5·85	4·55	5·54	5·35	5·74	5·19	6·01	5·59
Physics	5·56	4·03	5·08	4·62	5·34	3·95	5·40	4·27

[a] $a + c_i$ (see p. 38)

A full list of Schools Council working papers, curriculum and examinations bulletins and other such publications is available from the Information Section, Schools Council, 160 Great Portland Street, London W1N 6LL.

This series of examinations bulletins consists mainly of accounts of experimental and trial examinations for C S E, and their results. Many of the techniques described, however, are applicable to examining in general. The series also presents new thinking and proposals in the field of examinations.

Examining boards, universities, schools and H M Inspectorate have co-operated with the Schools Council in these studies, and in many cases the initiative has come from these sources.

The first four titles were published for the Secondary School Examinations Council whose work was taken over by the Schools Council.

Price £1·30 net