The Irish Journal of Education, 1969, III, 1, pp 5-21

MARKER RELIABILITY IN THE IRISH LEAVING CERTIFICATE*

JOHN MACNAMARA
Educational Research Centre, St Patrick's College, Dublin
and
GEORGE F MADAUS
Boston College

The reliability in marking of nine subjects in the Irish Leaving Certificate examination taken at the end of secondary (grammar) schooling was investigated Forty scripts in each subject were marked by two different examiners and by one examiner on two different occasions A high degree of unreliability was found in the marking of all subjects The sources of the unreliability are discussed and, in the light of the relevant literature, two principal ways for counteracting such unreliability are suggested the use of multiple-choice questions and the multiple marking of essays

At the end of secondary school (grammar school) Irish students sit for the Leaving Certificate Examination (LCE)—a public examination run by the government's Department of Education To a very great extent a student's future career depends on the results which he then obtains Admission to university, to other forms of third level education (e g teacher training) and to a wide range of employment (e g civil service and the bank) depends heavily on LCE marks † For example, in order to be admitted to university on LCE results a candidate must gain passes in at least five subjects, at least two of which must be at the honours level In addition, candidates who gain four LCE honours are eligible for a

*This study was financed by the Department of Education of the Irish Government The authors wish to thank the many persons in that Department who assisted in the study, especially Mr Sean O Connor and Mr Seamus Ó Ciarnáin The authors are also grateful to Mr Aidan Moran, then a statistician in An Foras Taluntais and now in the Department of Statistics at Trinity College, Dublin, for advice about the statistical analysis The present paper is based on a report submitted to the Department of Education (cf 17)

†See Madaus and Macnamara (17) The National University of Ireland still runs its own matriculation examination, but admission to the National University is almost exclusively by means of the LCE (cf 15, pp 371-373)

university scholarship (subject to a means test) The numbers who sit for the LCE are increasing rapidly, in 1967 they reached about 13,600 which is about 23 per cent of an age cohort (17) Thus for a large number of students, for their parents, for those who are engaged in the selection of students either for higher education or employment, and for the community at large the accuracy of LCE marks should be a matter of serious interest

The accuracy of examination marks is not a simple notion, but rather one which on closer study reveals several distinct aspects The first distinction is between validity and reliability A test is *valid* to the extent that it measures what it is supposed to measure Thus, a test of Latin composition which examines English vocabulary rather than knowledge of Latin would not, in spite of its title, be a valid test A test is *reliable* to the extent that the marks obtained with it are free from random error, i e to the extent that the marks reflect some characteristic of those to whom the test is applied rather than mere chance factors Clearly, reliability is a necessary but not sufficient prerequisite for validity Marks awarded on a totally unreliable test are simply random figures, they cannot measure anything validly However, a test can be both highly reliable and highly invalid Thus a highly reliable test of ability to *write* French might not measure ability to *speak* French at all If such a test were employed for the purpose of assessing ability to speak French it would be at once highly reliable and highly invalid

Reliability, which alone concerns us in this paper, is also a complex concept which needs to be analysed There are three major sources of unreliability in a test the questions contained in the test, the student, and the marker Invariably, the questions on an exam paper are a small sample of all the questions which might reasonably have appeared on the paper The only value of the ones which do appear lies in their representativeness—i e how accurately they represent the total body of possible questions To the extent that they are unrepresentative the test is not reliable Secondly, students can vary from one occasion to the next, they can have headaches, cramps, emotional upsets of various sorts, and they can have been lucky or unlucky in these respects on the occasion when they were tested We shall not be further concerned in the present paper with either of these sources of unreliability We shall confine our attention to marker reliability by which we mean (i) the extent to which *two different markers* agree in the marks they award to a single set of answers, and (ii) the extent to which *the same marker* on two different occasions is consistent in marking the same set of answers

METHOD

Our investigation was of the marking of nine LCE subjects in 1967 In all there were twenty-five LCE subjects from which we chose English, Irish, French, Latin, history, geography, mathematics, physics, and chemistry * In each of these subjects there were separate examinations for pass and honours candidates, which makes a total of eighteen examinations In each subject honours were awarded to candidates who obtained sixty per cent of the marks in the honours examination, passes were awarded to candidates who gained at least thirty per cent on the honours paper or forty per cent on the pass one

For the investigation of marker reliability a sample of forty answer papers was selected in each of the nine subjects In each subject separate samples of pass and honours papers were drawn The selection was made on the basis of a preliminary reading of the answer papers before they had been marked by the inspectors in charge of the markings, and the scripts were selected so as to represent a wide range of ability † Each script was then photocopied twice The originals were sent to official LCE markers who marked them in the usual way under the supervision of the inspectors the marks awarded them on this occasion were the official LCE marks At the same time the first set of copies was sent to an alternative group of markers who, working independently of the first group, also marked them The second group of markers were official LCE markers and they too worked under the supervision of the inspectors The second set of copies was kept until after Christmas when they were re-marked by the persons who had marked the originals We assumed that by then they would have forgotten both the papers and the marks they had previously given We were fortunate that although they had not been aware that they would be requested to do so, all these markers agreed to carry out the re-marking We thus obtained three sets of marks for almost a complete set of 720 scripts, i e 9 (subjects) $\times$2 (pass and honours) $\times$40 students Further, the marks for each answer in each script were recorded separately, and so we were able to study marker reliability in relation to each question Understandably in so large an undertaking a small number of copies were misplaced or spoiled For instance, the photocopying process failed to produce adequate copies of a few students'

*The subjects which we did not study are Greek, German, Italian, Spanish, Hebrew, applied mathematics, music, general science, botany, physiology and hygiene, physics and chemistry, agricultural science, domestic science, commerce, drawing and art

†Since there were no IQs available this seemed the appropriate way to proceed

graphs Nevertheless, the three sets of marks are complete for almost all the material

ANALYSIS

Two limitations of the study must be understood Only forty scripts for each examination were re-marked, and there are no grounds for believing that the selected scripts form a fully representative sample of all the scripts submitted in any particular examination However, we employed a form of statistical analysis for which the latter was not a prerequisite

The second limitation is more serious but could not have been overcome without incurring prohibitive expense For any individual examination we studied the marking of only two persons, and we know nothing of how these two persons compare with other persons who marked the remaining papers Similarly, the double set of marks obtained from a single marker do not enable us to infer with confidence how consistent other markers would have been if they had marked and re-marked a similar set of papers However, there was not much variation in marker reliability across the nineteen examinations which were investigated, and it is not unreasonable to suppose that the outcome would have been similar if a different set of markers had been selected

Perhaps the measure of reliability most frequently employed is the Pearson product-moment correlation coefficient * Underlying the use of this coefficient is the assumption that one is dealing with a normal bivariate distribution There is little reason to suppose, however, that the ability or abilities measured in any of the LC exams, pass or honours, was normally distributed among the candidates who took it Moreover, we have just seen that there is no reason to believe that the forty scripts for any exam were representative of the entire body of scripts for that exam Consequently, the Pearson product-moment correlation coefficient is hardly appropriate Instead we used a simpler and more direct measure, not of reliability but of unreliability Before describing this measure, however, it is necessary to discuss one or two preliminaries

Error associated with the marks awarded by a single individual can be resolved into two components, (i) general bias and (ii) random fluctuation

*Readers who are unfamiliar with the technical details of this and the succeeding paragraphs will find explanations in McNemar (18, chapters 8, 9 and 10), for example—or they may simply skip this section and go straight on to the beginning of the next section

The first arises from the individual's tendency to mark 'hard' or 'easy' relative to the 'average' marker Ideally, the correct estimate of general bias would be the difference between the average of a set of marks assigned by a particular marker and the average marks assigned by the entire group of markers to the same set of papers Needless to add, we could not obtain this estimate since we did not have all markers assign marks to the same set of papers The only evidence available to us that general bias exists in the marking of the LCE is furnished by the difference between the average marks assigned either by two different markers or by the same marker on two different occasions The relevant data—mean differences—are recorded in Table 1 in the columns headed $\overline{D}$

The second component of error, random fluctuation, can be attributed to simple inconsistency To express the idea unkindly, this means that the examiner's bias does not operate consistently, there is a degree of random error associated with how his bias operates Statistically, to treat marker-error as due to these two sources is equivalent to interpreting it as composed of a constant difference between each pair of marks in two sets of marks (general bias) and a remainder which contains all further inconsistency (random fluctuation) *

Our measures of random fluctuation were calculated by subtracting either the second reader's marks from those of the first, or a reader's second set of marks from his first, and calculating the standard deviation (*SE*) of the resulting differences On the assumption that such differences are normally distributed about their mean we can enter tables of normal distribution with the calculated *SD*s and estimate confidence intervals for LCE marks

The assumption that difference scores of this type are normally distributed is quite reasonable over most of the range of marks in most subjects Where it might at first sight appear to be untenable is in relation to marks close to either zero or to the maximum mark An answer which has received a mark of zero from one reader can receive zero or higher from a second reader, it cannot receive a lower mark Similarly, an answer which has received the maximum mark from one reader can receive the same or a lower mark from another reader, it cannot receive a higher mark Thinking that very low and very high marks would be particularly common in mathematics, we paid special attention to marks in that subject We plotted scattergrams of the marks for each question in mathematics but observed no tendency for difference scores to be

*The statistically knowledgeable reader will see the relationship between this model and the one used in analysis of variance

TABLE 1

ESTIMATES OF MARKER UNRELIABILITY

	PASS						HONOURS						
	Two markers			Single marker			Two markers			Single marker			Maximum
	$\bar{D}$	*SD*	N	$\bar{D}$	*SD*	N	$\bar{D}$	*SD*	N	$\bar{D}$	*SD*	N	mark
English	41 5	23 0	39	4 7	19 3	40	2 2	29 6	40	—3 6	13 6	40	400
Irish	4 0	23 6	40	1 8	14 4	40	4 1	2 7	40	17 0	37 1	40	500
French	25 7	18 7	33	18 8	11 8	38	37 0	14 8	40	—4 1	15 5	37	400
Latin	—3 1	18 2	40	—3 9	8 8	40	—32 3	14 6	39	3 4	14 2	39	400
History	12 5	14 8	40	—12 7	11 2	40	6 7	13 5	40	—5 7	16 4	37	300
Geography	10 5	20 4	39	2 8	15 6	40	7 8	15 2	38	—0 7	4 2	40	300
Mathematics I	2 0	11 2	40	2 5	10 4	40	1 4	14 6	40	9 7	15 7	40	300
Mathematics II	—11 1	13 9	40	0 4	15 0	40	18 6	14 4	40	—5 3	14 9	40	300
Physics	—39 0	18 8	40	8 1	14 2	40	14 9	20 9	40	—18 2	20 5	40	400
Chemistry	8 1	11 2	40	3 8	9 7	40	—11 6	19 0	40	—6 3	18 9	38	400

smaller towards the extremes This is probably due to the fact that each mathematics question is composed of several subsections which are marked separately, while we had before us only the marks for the question as a whole In conclusion, then, the assumption that difference scores are normally distributed is quite in order

RESULTS

In Table 1 where the results are laid out there are two sets of figures associated with mathematics In pass and honours mathematics there are two papers each marked out of 300 In each exam the scripts for the first and second papers were written by different students, so the data for the two papers cannot be combined

The way to interpret Table 1 can best be explained by taking an example, say, pass English The mean difference between the marks assigned by the two different readers is 41 5 ($\overline{D}$), the standard deviation of difference scores is 23 0 (SD) Since the marks assigned by the second reader were (throughout the table) subtracted from those assigned by the first one, $\overline{D}$=41 5 indicates that the second reader was 'harder' than the first one by that number of marks on an average (i e 10%) The SD=23 0 indicates (by means of the table of normal probability) that due to random fluctuation eight students in twenty-five would be expected to receive from the second reader totals differing by at least 23 marks from those they received from the first one Further, one student in twenty would be expected to receive from the second reader a total differing by at least twice that amount (46 marks) from the total he received from the first reader However, error due to general bias (D) and error due to random fluctuation (SD) must be combined if we are to reach a realistic appreciation of the total error involved For one student in twenty this can be calculated from the formula observed total $-\overline{D}\pm 2SD$ For example, if the first reader allotted a pass English student a total of 200 (50%) the chances are one in twenty that the second reader would allot him a total of 200—41 5+46=204 5 (51%) or more, or 200—41 5—46=112 5 (28%) or less The latter mark is a fail The corresponding figures for marks assigned by a single reader on two different occasions are 233 9 (59%) and 156 7 (39%) The latter mark falls just short of a pass

A glance through Table 1 as a whole reveals that the unreliability associated with the marks assigned by a single reader on two different occasions is scarcely lower than that associated with the marks of two different readers This finding corroborates those of numerous other

investigations (cf 33, p 205) Further, throughout the table as a whole there is generally a one in twenty chance for marks to swing up and down by about ten per cent in each direction These figures apply to the totals for individual subjects, if subject totals were combined to yield an overall assessment of a student, the percentage fluctuation in marks would be less The errors in individual totals would tend to cancel each other However, LCE results for different subjects are generally treated separately, and so the figures given in Table 1 are the appropriate ones with which to measure marker reliability

DISCUSSION

It is important to appreciate that the extent of marker unreliability in the LCE is not due to any carelessness on the part of examiners or markers Such unreliability has everywhere been found in association with essay-type examinations similar to the LCE Indeed the Department of Education takes numerous precautions to guard against bias and error Candidates, names are withheld from markers, schematic answers to each question are prepared with detailed marking instructions, a conference of markers is convened to discuss difficulties which are likely to arise in the marking and several examiners supervise the marking, each examiner being responsible for maintaining a uniform standard among a group of markers The observed unreliability, then, arises despite conscientious efforts to avoid it

A closer look at the marker's task and at the marking directions he receives reveals the source of the trouble For example, the total of 120 marks for the English essay, an item notoriously troublesome to markers, was broken down into a certain number for ideas, a certain number for expression, and a certain number for English usage On the other hand the instructions contain the general directive

> Read the instructions given in the marking scheme, and keep them in mind However, in practice, you will usually find it more satisfactory to mark the composition on the basis of your general impression of its worth

Apart from the apparent contradiction in these instructions it is immediately evident that much is left to the taste, judgment and discretion of the individual marker When so much must be left to the individual marker, it is hardly surprising that marker unreliability should be as high as it appears to be

While many persons will not be surprised to learn that there is considerable unreliability in the marking of essays, they may be surprised to learn that marker unreliability in mathematics is scarcely less pronounced The reason can again be traced to the task which the markers were expected to perform For example, the marking scheme for the second pass paper starts with the following instructions

> Blunders or serious omissions (—10) each A very serious blunder on an item may entail the loss of all marks for that item Numerical slips (—3) each

Further down the same marking scheme states 'Less serious blunders or omissions (—5) each' Thus this single marking scheme refers to four types of mistakes slips (—3), less serious blunders or omissions (—5), blunders or serious omissions (—10), a very serious blunder (loss of all marks) Marker unreliability, of course, is directly traceable to the lack of a definition of any of the types of mistake and to the lack of precise criteria for distinguishing between them The marking scheme for the first honours paper does give a small number of unambiguous instructions about the treatment of specified mistakes, but apart from these the marking scheme is not nearly precise enough to preclude very wide variation in marking Furthermore, the practice of awarding marks for an 'attempt' is likely to lead to unreliability For example, the marking scheme for question 7 on the second pass paper says that an attempt is worth 30, or 20 or 10 marks, but no directions are given for distinguishing between the three levels Moreover the relationship between attempts, on the one hand, and slips, blunders and omissions on the other hand is not specified

No doubt, at the markers' conference prior to marking serious efforts were made to elucidate some of these and numerous other obscure points However, it is unlikely that an even remotely satisfactory clarification could have been reached Indeed, it is probably impossible in principle to be sufficiently explicit since while there may be only one right answer, the possibilities for wrong answers and for errors are indefinitely large

Other less obvious sources of marker unreliability are well known to specialists in testing One such is operative when readers mark several questions in succession from the same individual In marking later answers readers tend to be biased by what they have already seen of a candidate's work Further, the evaluation of essay-type answers has frequently been found to vary with the candidate's literary style, grammar, spelling and penmanship, even when markers have been explicitly warned to pay no

attention to them Nyberg (21) factor-analysed essay marks assigned to high-school leavers in the Province of Alberta, Canada, and found that the mechanics of English writing—spelling, punctuation, word usage, grammar—contributed more to the overall mark than any style-content factors Spelling alone contributed more to the overall mark than the combined style-content variables such as vividness, relevance, originality and organization Further Rothkopf and Turner (29) have shown that essays which employ a technical vocabulary tend to gain higher marks than similar essays which do not

The practice of allowing optional questions gives rise to marker error in several ways One of the assumptions underlying this practice is that the marking of all answers is equally lenient or equally severe However, the evidence is clear (6, 20, 31) that irrelevant factors affect the judgment of markers and invalidate the assumption For instance, markers tend to be stricter in their assessment of a question which has been attempted by many persons than in their assessment of one which has been attempted by only a few Moreover, if most students answer a particular question well, markers become more and more severe as they read through the responses

RECOMMENDATIONS

We do not wish to make a fetish of reliability, we realize that it is secondary to such matters as the abilities and information demanded of students by examiners and the effects of these demands on the whole educational process Nevertheless, there is a serious question of justice involved, so we feel it incumbent upon us to seek out remedies for the high level of marker unreliability in the LCE It will be a matter of concern, however, that our suggestions for reliability will not be at the expense of validity, and above all we shall try to ensure that our suggestions will not, if carried out, have a detrimental effect on the education of secondary school students

The simplest method of reducing marker unreliability is to employ multiple-choice questions Essentially such questions differ from the essay-type questions in that several responses are suggested to the student who is required to indicate what he considers to be the correct one His response can then be assessed by reference to a very explicit marking scheme—the student is right if he has chosen one particular response and wrong if he has chosen any other one For example, candidates for honours English were required to define the word *formidable* as used in a given

prose passage In multiple-choice form the item might have read something like this

> Select from among the alternatives marked A to D the word which comes closest to the meaning of the underlined word as used in the preceding passage—
>
formidable	A	enormous	B	frightening
	C	ingenious	D	unmanageable

Responses to such questions can be checked with absolute accuracy by a mechanical device known as an optical scanner The use of such a scanner would remove from examining much of the present drudgery and leave markers free for other tasks Furthermore, contrary to popular opinion it is possible to frame multiple-choice items which test a wide range of intellectual abilities (cf 2, *passim*) The classical source on the writing of such items is E F Lindquist's *Educational measurement* (16), further help will be found in Adkins *et al* (1), Ebel (7), Furst (10), Gronlund (13) and in a less theoretically orientated book by Gerberich (11)

Few topics are more likely to disturb the equanimity of European educationalists than multiple-choice examinations Fears are aroused partly due to the unfamiliarity of such examinations, partly due to the belief that they can evaluate factual information only, partly due to the belief that they lead to a fragmented approach to education and partly to the belief that they attempt to render mechanical and impersonal what is essentially human and personal All this amounts to the fear that multiple-choice items will rapidly destroy all that is best and wisest in the tradition of European education And truth to tell there is a certain distain for a type of examination which is closely associated in people's minds with America

The last observation apart, we are to a great extent sympathetic to the fears of European educators which are shared by many Americans too, and we are fully aware that between examinations which are detrimental and those which are beneficial in their effects runs a faint and delicate line But we must be clear that the mischief so readily attributed to multiple-choice examinations may just as easily result from traditional ones (17 *passim*)

Objections to multiple-choice examinations vary widely and some are better based than others For instance many teachers who are un-

accustomed to such exams fear that by chance alone some candidates will give the correct answer to a large number of questions It is of course true that by guessing candidates can give a certain number of correct responses, but a candidate who answers correctly only one-fourth or one-fifth (depending on the number of alternatives) of the questions in a multiple-choice paper is treated as roughly the equivalent of a candidate who obtained a mark of zero in a traditional examination Neither has given any evidence of learning or of intellectual ability In other words the fear that success and failure in such exams is *largely* a matter of chance is ill-founded, the likelihood of unfair results is real but far lower than in traditional exams For full discussions of the effects of guessing the reader is referred to Davis (5) and Traxler (32)

A more serious, and more solidly based, criticism of multiple-choice questions is that answers to them are generally marked either right or wrong If right, an answer gains one mark, if wrong, it gains nothing Contrary to commonsense, the method treats all wrong responses as equally wrong, and, by assigning a constant mark for a correct response, it treats all correct responses as equally correct and deserving of marks In order to remove the grounds for this legitimate complaint Ramsay (25) has recently proposed a method of assigning weights to each of the alternative answers associated with a multiple-choice question The weights would have the effect of rewarding some correct responses more than others and also that of penalizing some wrong responses more than others The calculation of weights depends on the prior identification of two groups of candidates, a bright one and a dull one The division into groups could be made by any suitable or convenient means e g teachers' estimates, or the total number of correct answers given by individual candidates to the test questions Once the two groups are established, there are mathematical procedures which will yield weights of the type described for the purpose of maximizing the test's power to discriminate between the two groups These weights would have the added advantage of reducing to zero the expected mark for a person who randomly guessed the answer to each question Other approaches to the same problem have also been suggested (28, 30)

However, the most serious limitation on the use of multiple-choice questions is the assumption that there is a best, or correct, response to each question, and that the correct response can be determined without consulting candidates' responses Generally speaking the assumption causes little difficulty in school mathematics, physics and chemistry, and in questions in any subject which merely call for a detail of information

The assumption is likely to cause difficulty in connection with higher order questions about literature or history, e g questions about the overall evaluation of a poem or of a historical personage's character Partly in view of this limitation we advocate the retention of essay questions (see below) in certain areas it is wisest to have candidates put their case and support it with their own arguments However, this is not to grant that multiple-choice items are unsuitable for assessing higher-order abilities in literature or history It is merely to allow that such items should not be employed where a best answer cannot be determined and where the purpose of the exercise is to elicit candidates' argued responses and then evaluate them for such logical qualities as organization, comprehensiveness, coherence, and reasonableness of ideas, and for such literary qualities as sensitivity to language and lucidity of expression In most cases, however, it is possible so to select alternative responses to a multiple-choice question that authorities would agree on a best response, and in that case a multiple-choice item is appropriate Moreover, the difficulty we are discussing is by no means confined to multiple-choice exams Though the examiner who has set an essay-type exam may have decided that there is no best answer, the markers who read the answers are likely to have preferences which influence their assessment of individual candidates It seems likely then that the problem of best, or correct, answers is quite as formidable, though more treacherous being less obvious, in essay exams All things considered, the argument for the extensive use of multiple-choice questions is very strong

We fully agree, however, that an examination composed exclusively of multiple-choice items would fail to assess adequately certain qualities which are more conveniently (though never certainly or easily) assessed in an essay—originality of approach, organization of ideas and style of writing (9) We do not propose that essay questions should be excluded from the LCE We simply suggest that essay questions should be employed only for the purpose for which they are best suited, for the rest we suggest that multiple-choice items be employed Incidentally, in almost all the LC examinations which we studied a very large proportion of the marks went for the recall of details of information Clearly such objectives are best assessed by means of the multiple-choice item However, both from the point of view of the examination and from that of the effect of the examination on education it is highly desirable that the papers for each subject should, where possible, carry a certain number of essay questions The problem, then, is to reduce drastically the unreliability to which the marking of essays is subject

The many attempts in the past to improve the reliability of essay marks have for the most part proved disappointing and serve as a warning that the task is a difficult one For example, Noyes, Sale and Stalnaker (cited in 31) found that despite the preparation of a carefully defined marking scheme which markers were assiduously trained to apply, the coefficient of marker reliability rose no higher than 58 This figure means that the marks assigned to essays were practically devoid of any usefulness

Rippey (26, 27) has had greater success (coefficients of 80 and above) by intensively training readers to use a scheme in which certain skills, qualities and components of writing ability are painstakingly defined and illustrated with samples of students' writing His procedures are extremely costly in time and money, however, and are thus hardly suitable for use with large numbers of scripts

The most hopeful line of approach is one which was explored twenty years ago with children's essays by Wiseman (34), Finlayson (8) and others, and recently taken up again by Pilliner (23, 24) and especially by Coffman and his colleagues at Educational Testing Service Coffman and his group have also had considerable success in the use of indirect measures of writing skill Godshalk, Swineford and Coffman (12), for example, compared the relative merits of three methods of assessing writing skill The first was called 'objective items' and consisted of a series of multiple-choice questions grouped under the following headings usage, sentence correction, paragraph organization, prose groups, error recognition, and construction shift The second, 'interlinear exercises,' need not detain us here In the third, called 'essay exercises,' students were asked to write three paragraphs and two longer essays all on different topics twenty minutes were allowed for each paragraph and forty minutes for each essay Five readers independently assigned one of three marks to each of a candidate's five responses to the essay tasks '3' for a superior response, '2' for an average response, and '1' for an inferior one Readers based their marks on global impressions rather than on an analysis of the writing The marks of the five readers for any particular piece of writing were added to yield totals ranging from 5 to 15 marks, and each student's five totals were then combined to yield a 'composite' (or average) score for his total performance The results were, (i) the reliability of essay marks is primarily a function of the number of essays written by each student and of the number of readers The rehability coefficient for composite scores (five essays and five markers) was 92, the reliability coefficient of a single marker's assessment of a single essay was 40 These figures closely match those obtained by Wiseman, Finlayson, Pilliner and others (ii) The

objective items proved to be reasonably valid; they yielded correlations ranging from .56 to .70 with composite essay marks. Further work with the objective items would be necessary, however, before they could be recommended for general use.

The advantage of the global over the analytic approach to essay marking is borne out in a study done by Coffman and Kurfman (3). They found that the global approach yielded substantially higher reliability coefficients. In concluding this study the authors point out that the possibility of marker unreliability should be demonstrated to markers, and they suggest that daily conferences of markers should be held to discuss marks assigned by different markers to sample essays. In this way it is possible to obviate the criticism by Cox (4) that increases in marker reliability of the type we have been discussing may arise less from agreement among different markers about the value of essays than from each marker's stubbornly holding to some private standard.* Perhaps the most encouraging result of all is the finding by Myers, Coffman and McConville (19) that the level of reliability obtained with the Coffman approach held up over a period of five days during which 80,000 essays were marked.

The Coffman technique is certainly more expensive than that used in marking the LCE at present, but not prohibitively so. If our suggestion to conduct a large part of the examinations by means of multiple-choice items were adopted, the marker's work would be greatly reduced. The marking of the multiple-choice sections could be done cheaply by means of an optical scanner, and markers would be free to do that which they alone can do, mark the essays. In view of the importance of the results to students and the importance of good examinations to education we feel that the added expense involved in the Wiseman-Coffman technique would be amply justified.

REFERENCES

1. Adkins, D. C. *et al. Construction and analysis of achievement tests.* Washington, D.C.: US Civil Service Commission, 1947.

2. Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. *Taxonomy of educational objectives. Handbook 1: Cognitive domain.* London: Longmans, 1956.

3. Coffman, W. E., and Kurfman, D. A comparison of two methods of reading essay examinations. *American Educational Research Journal,* 1968, *5,* 99-108.

*However, see Pilliner (24) for a strong statistical argument to the effect that Cox is unlikely to be right and that marker reliability obtained by increasing the number of markers is likely to derive from capitalizing on the small measure of agreement that invariably exists between markers about the relative merits of a group of essays.

4 Cox, R Examinations and higher education A survey of the literature *Universities Quarterly*, 1967, *21*, 292-340

5 DAVIS, F B Item selection techniques In Lindquist, E F (Ed), *Educational measurement* Washington, D C American Council on Education, 1951 Pp 266-328

6 DEVADASON, M D Optional questions in tests and examinations *Teacher Education*, 1963, *8*, 63-67

7 EBEL, R L *Measuring educational achievement* Englewood Cliffs, N J Prentice-Hall, 1965

8 FINLAYSON, D S The reliability of marking of essays *British Journal of Educational Psychology*, 1951, *21*, 126-134

9 FRENCH, J W Schools of thought in judging excellence of English themes *Proceedings of the 1961 Invitational Conference on Testing Problems* Princeton, N J Educational Testing Service, 1962 Pp 19-28

10 FURST, E J *Constructing evaluation instruments* New York David McKay 1958

11 GERBERICH, J R *Specimen objective test items A guide to achievement test construction* London Longmans, Green, 1956

12 GODSHALK, F I, SWINEFORD, F, and COFFMAN, W E *The measurement of writing ability* New York College Entrance Examination Board, 1966

13 GRONLUND, N E *Measurement and evaluation in teaching* New York Macmillan, 1965

14 *Investment in education* Report of the Survey Team appointed by the Minister for Education in October 1962 Dublin Stationery Office, 1966

15 *Investment in education* Annexes and Appendices to the Report of the Survey Team appointed by the Minister for Education in October 1962 Dublin Stationery Office, 1966

16 LINDQUIST, E F (Ed) *Educational measurement* Washington, D C American Council on Education, 1951

17 MADAUS, G, and MACNAMARA, J *Improving the public examination A Study of the Irish Leaving Certificate* Report submitted to the Department of Education Dublin Educational Research Centre, St Patrick's College Unpublished manuscript, 1969

18 MCNEMAR, Q *Psychological statistics* (2nd edition) New York Wiley, 1962

19 MYERS, A, MCCONVILLE, C B, and COFFMAN, W E Simplex structure in the grading of essay tests *Educational and Psychological Measurement*, 1966, *26*, 41-54

20 NATIONAL COUNCIL OF EDUCATIONAL RESEARCH AND TRAINING *Optional questions in tests and examinations* New Delhi NCERT, 1963

21 NYBERG, V R The reliability of essay grading Paper read to Sixth Canadian Conference on Educational Research Ottawa Canadian Council for Research in Education, 1968 Mimeographed

22 PILLINER, A E G The application of analysis of variance to problems of correlation *British Journal of Psychology (Statistical)*, 1952, *5*, 31-38

23 PILLINER, A E G Examinations In Butcher, H J, and Pont, H B (Eds), *Educational research in Britain* London University of London Press, 1968 Pp 167-184

24 PILLINER, A E G Multiple marking Wiseman or Cox? *British Journal of Educational Psychology*, in press

25 RAMSAY, J O A scoring system for multiple-choice test items *British Journal of Mathematical and Statistical Psychology*, 1968, *21*, 247-250

26 RIPPEY, R Criterion referenced tests for English compositions Paper delivered to the National Council for Measurement in Education Chicago Center for the Cooperative Study of Instruction, University of Chicago, 1965 Mimeographed

27 RIPPEY, R Cognitive maps for English composition Paper read to the American Educational Research Association Chicago Center for the Cooperative Study of Instruction, University of Chicago, 1966 Mimeographed

28 RIPPEY, R Probabilistic testing *Journal of Educational Measurement*, 1968, *5*, 211-215

29 ROTHKOPF, E Z, and THURNER, R D Effects of written instructional material on the statistical structure of test essays Murry Hill, N J Bell Telephone Laboratories, 1968 Mimeographed

30 SHUFORD, E, ALBERT, A, and MASSENGILL, H Admissable probability measurement procedures *Psychometrika*, 1966, *31*, 125-145

31 STALNAKER, J M The essay type of examination In Lindquist, E F (Ed), *Educational measurement* Washington, D C American Council on Education, 1951 Pp 495-530

32 TRAXLER, A E Administering and scoring the objective test In Lindquist, E F (Ed), *Educational measurement* Washington, D C American Council on Education, 1951 Pp 329-416

33 VERNON, P E *The measurement of abilities* London University of London Press, 1940

34 WISEMAN, S The marking of English compositions for grammar school selection *British Journal of Educational Psychology*, 1949, *19*, 200-209