The Irish Journal of Education 1974 *viii* 1 *pp* 36-48

EVIDENCE FOR DIFFERENTIAL MARKING DISCRIMINATION AMONG EXAMINERS OF ENGLISH

Bob Wood and Douglas Wilson*
University of London School Examinations Department

School examinations are undergoing changes but one feature which is unlikely to alter greatly is the marking of extended writing Being dependent upon examiners, differences are liable to occur which, until recently went largely unrecognised In this study a University of London GCE O level English Language essay paper was used to investigate examiners' marking behaviour in particular the way they used the mark scale The use of a multiple-choice comprehension paper mark as a concomitant observation suggested that examiners were discriminating in a non uniform manner between candidates

INTRODUCTION

There has been some research on determining the traits examiners look for and reward and the external factors influencing their marking (see, for instance, 4, 3, 7), but little has been published on the routine monitoring of the marking behaviour of examiners marking a large number of candidates

Where the number of candidates is measured in tens of thousands, the precision of examinations is dependent upon a stable marking process Marks are awarded by a group of examiners who usually read between 200 and 500 scripts apiece

The University of London Examinations Department uses an analytical marking procedure which it is hoped curbs examiner differences although the impossibility of eliminating subjectivity is always liable to lead to examiner differences in level and precision of marking From the examining board's point of view it is desirable to be able to detect the form that these differences take and, if possible, to take more or less immediate corrective action Ideally comparison between examiners can best be made by requiring all scripts to be marked by each examiner in turn An analysis might then proceed along the lines described by Cochran (2) However, for an operational examination this method is obviously impractical in terms of time

* Requests for off prints should be sent to R Wood University of London School Examinations Department 66 72 Gower Street, London, WC1E 6EE

and expenditure, although some information can be gained from having examiners mark photocopies of a sample of scripts at the outset, a practice which is routine in most examination boards However, doubts are always expressed about the typicality of photocopy marking and about the assumption that marking behaviour will remain stable during the marking stint Indeed evidence of our own suggests it does not For these reasons we have concentrated on learning what we can from the whole of an examiner's work

A big problem in studying examiner differences in the operational situation is that examiners are not allocated scripts of the same quality (see below) Some means of compensating for this must be introduced Suppose that, in addition to the written paper, candidates also take a paper which can be marked in a perfectly standardised way, a multiple-choice paper for instance This concomitant information can now be utilised in a covariance analysis to investigate examiner sample differences

Other factors that are likely to affect the analysis are age and sex of candidate and centre type Statistics show notable differences in the proportions of boys and girls passing O level GCE subjects, with girls enjoying higher pass rates, and English Language is no exception Thus it is certainly necessary to allow for differing sex ratios in samples of scripts The age of candidates is more or less completely confounded with type of centre or school and it has been left out of the analysis Centre type is likely to be a more serious source of variance, because the scripts marked by examiners usually come from a limited number of centres and the method of allocating scripts to examiners does not, unfortunately, make allowances for centre comparisons on a within examiner basis

DATA

The London Board 1973 O-level English Language examination consisted of two papers, a multiple-choice paper (MCP) giving a maximum possible mark of 60, and a written paper (WP) comprising essay and summary components with a maximum possible mark of 65 The number of candidates sitting the examination was in excess of 50,000 but of these only 31,436 records were in a form available for analysis The examiners who marked the available scripts numbered 72

The rationale of allocation is based on the proposition, palpably untrue in places, that the spread of ability within a centre can be taken as being representative of the population In fact, three examiners marked only female candidates The extent of violation of the assumption can be gauged from the histograms of MCP means shown in Figure 1 The skewness co

efficient $\sqrt{\beta_1}$ was generally in the region 0 20 and kurtosis β_2 was approx imately 2 90 In one or two samples kurtosis was as high as 4 with marked skewness but this was to be expected as some schools are selective and the ability of their candidates would be above average

The corresponding histograms for the WP means are given in Figure 2 The skewness coefficient was generally somewhat higher for the written paper, being more in the region of 0 30 From Figure 1 and the values of $\sqrt{\beta_1}$ and β_2 it can be seen that only 3 examiners received script samples which were wholly unrepresentative of the population While their marking behaviour might well differ from that of the other examiners due to the high or low ability levels of their sample, (see Figure 3), it does mean that the basic condition for a meaningful covariance analysis – parity of groups on the independent variable – is more or less satisfied

ANALYSIS

The analysis is concerned with the following measures
examiner WP means i e mean score awarded to the candidates by the examiner,
examiner WP means adjusted for MCP mark,
examiner WP means adjusted for sex of candidates and MCP mark

In an obvious notation, the expected mean, m_i, for examiner i, without any adjustment, can be written as

$$m_i = w_i$$

An even simpler model, assuming that a constant WP mean were to be fitted, would be

$$m_i = w$$

When an adjustment for the MCP mark, x, is made the expression for the expected mean becomes either

$$m_i = w_i + bx$$

or

$$m_i = w_i + b_i x$$

depending on whether uniform regression or discrimination across examiners is assumed

The allowance for sex takes the form of a constant S which is meant to

account for the differences in performance of boys and girls, so that

$$m_1 = S + w_1 + b_1 x$$

where w_1 is now the mean adjusted mark for boys

We have been referring to the expected mean but, of course, we are assuming each awarded mark contains an error term, distributed according to the usual assumptions To estimate the consistency of an examiner's marking, the appropriate residual mean square, v^2, being a variance estimate, serves as a measure Multiplying assumptions, this term could be homoscedastic, $v^2 = \delta^2$ for all examiners, or heteroscedastic, $v^2_1 = \delta^2_1$

Thus it can be seen that the model fitted for any examiner has two terms The first is a regression of WP mark on MCP mark or some such linear function which can be homogeneous in all, some or none of its parameters over examiners The second term is an error component, with variance which can be homoscedastic or heteroscedastic across examiners To fit the appropriate model a generalised covariance analysis with a maximum likelihood solution was used, assuming the 'true' marks to be normally distirbuted The programme was based upon the method of Ashford and Brown (1), incorporating the simplex optimising routine of Nelder and Mead (5) and O'Neill (6)

The profusion of possible models was curbed immediately by considering examiners to be either all uniform for a parameter in any term or all different Models where some subgroup of examiners might have a term or parameter which was homogeneous for that group but not for another were ruled out This is because we are concerned with examiners as a whole and not with differences between any two examiners or groups of examiners Either all examiners are considered to be marking uniformly or they are not, the fact that some examiners may not differ from each other does not alter the conclusion that examiner marking behaviour varies

The consistency of examiner marking is not adequately reflected by the error variance estimate alone Joint consideration of the variance estimate, v^2_1, and the regression estimate b_1 is necessary An examiner's use of the mark range will be reflected in the size of the regression coefficient A small value will signal marking over a restricted range, whilst examiners with particularly high regression coefficients are marking distinctions between candidates of the same ability (as measured by MCP) that are generally not felt to exist by other examiners Because they are using the full mark range, their error variance estimates will inevitably be on the high side so that the ratio measure b_1/v_1 provides a better measure of an examiner's discriminating power

RESULTS

The complexity of the analysis makes comparison and interpretation of the different models a piecemeal operation The log likelihoods of the different models fitted are given in Table 1 Because of the heteroscedastic error term all tests used Bartlett's asymptotic chi square approximation for the null hypothesis of no difference between models

To aid comparison, the different models are divided into those with a heteroscedastic error term and those without, and in another, split into those allowing for sex differences and those not The test statistics for the various comparisons are given in Table 2 In all cases the models which specify heterogeneous regression and adjusted mean, numbers 4 and 8, have the best fit, for both types of error The likelihood ratio test for deciding whether the inclusion of a sex parameter is necessary gives a significant reduction in the likelihood, for the homoscedastic case the chi square value was 1236 with one degree of freedom, whilst for the heteroscedastic comparison the value was 800, also with one degree of freedom The test of whether heteroscedastic errors improve the fit had a chi square of 1042 with 71 degrees of freedom The estimated sex difference for this model was 2 32 with standard error 0 07

The range of the raw WP mean marks was 23 4 to 37 1, whilst for model 8 with heteroscedastic error the adjusted WP means ranged over the interval 4 5 to 25 7 The histogram for these two models (Figure 2) indicates that after adjusting for MCP marks the differences between examiner means are more noticeable The corresponding figures for examiner variances are 24 00 to 98 00 for raw marks and 19 66 to 60 75 for the adjusted marks given by model 8 with heteroscedastic error For this model the regression coefficients lie in the interval 0 042 to 0 628

The parameters b_1 and v^2_1 describing examiners' discriminating power are plotted in Figure 4 It can be seen that for nearly all values of the regression coefficient the error variance has a wide range of values Examiners who are equally discriminating between candidates as measured by b_1 are distinguishing differentially between candidates with the same MCP mark As explained earlier, a better measure of discriminating power is the ratio b_1/v_1 and the distribution of this index is drawn in Figure 5 The examiners who have index values outside the central range 0 055 to 0 085 and whose marking behaviour is considered to be wayward, are listed in Table 3

TABLE 1

LOG LIKELIHOODS OF MODELS

MODEL NUMBER	MEAN TERM	WITH HOMOSCEDASTIC ERROR TERM	DEGREES OF FREEDOM	WITH HETEROSCEDASTIC ERROR TERM	DEGREES OF FREEDOM
1	$w + bx$	101375	3	101104	74
2	$w + bx$	100039	74	99499	145
3	$w + b_1x$	100055	74	99519	145
4	$w + b_1x$	99745	145	99206	216
5	$w + S + bx$	100743	4	100404	75
6	$w + S + bx$	99443	75	99384	146
7	$w + S + b_1x$	99452	75	98992	146
8	$w + S + b_1x$	99127	146	98606	217

TABLE 2

ANALYSIS OF LOG LIKELIHOODS

SOURCES OF VARIATION	MODELS COMPARED	HOMOSCEDASTIC ERROR		HETEROSCEDASTIC ERROR	
		TEST STATISTIC	DEGREES OF FREEDOM	TEST	DEGREES OF FREEDOM
Heterogeneous adjusted means/ common regression	1 and 2	2672	71	3210	71
Heterogeneous regressions/ heterogeneous means	2 and 4	588	71	586	71
Heterogeneous regressions/ homogeneous means	1 and 3	2640	71	3170	71
Heterogeneous means/ heterogeneous regressions	3 and 4	620	71	626	71
Heterogeneous means/ homogeneous regression and sex	5 and 6	2582	71	2040	71
Heterogeneous regressions/ sex and heterogeneous means	6 and 8	632	71	1556	71
Heterogeneous regression/ sex and homogeneous means	5 and 7	2600	71	2824	71
Heterogeneous means/sex and heterogeneous regressions	7 and 8	650	71	772	71

TABLE 3

EXAMINERS WITH EXTREME VALUES OF DISCRIMINATION INDEX

ADJUSTED W P MEAN	REGRESSION COEFFICIENT	ERROR VARIANCE	INDEX
13 0	0 448	26 70	0 087
21 2	0 172	39 96	0 028
8 2	0 628	32 59	0 110
13 5	0 438	25 53	0 087
11 9	0 446	24 53	0 090
12 7	0 506	25 70	0 100
18 7	0 255	22 78	0 053
25 0	0 204	21 57	0 044
18 2	0 315	39 94	0 050
8 7	0 533	28 99	0 099
19 7	0 204	36 02	0 034
25 7	0 042	33 03	0 007
19 3	0 251	23 93	0 051
24 1	0 283	33 03	0 049
20 1	0 266	30 03	0 049
20 2	0 286	59 20	0 037
11 9	0 547	36 27	0 091

DISCUSSION

The log likelihoods of Table 1 demonstrate that there is no simple way of accounting for differences between examiners' marking behaviour We have found that the most unrestricted model has given the best result Although there are other sources of variation which could be taken into account and more complicated models constructed, differences between examiners' adjusted means would still be noticeable

The use of different regressions for the sexes appears unlikely to have much effect on examiner adjusted means A more likely source of variation is type of centre Some examiners will mark perhaps only Grammar School pupils, others private entries or Further Education candidates with a wide range of ability Context effects being what they are, over all calibre of scripts is bound to exert an effect on marking behaviour

Although the regression of written mark on multiple choice mark was used to accommodate sample differences, examiners evidently do not have a common measure for discriminating between candidates, even ignoring those examiners whose samples were peculiar to small numbers The rejection of those examiners (Table 3), who most obviously differ from the majority, would enable the remainder to be represented by a common regression coefficient somewhere in the region 0 35 to 0 45, although,

strictly, a more unrestricted model, with a *group* of examiners being represented by the same parameter value, is what is wanted A study of the results of model 8 for the homoscedastic and heteroscedastic cases indicates that for practical purposes of adjusting marks the existence of differential error can be ignored Figures 6 and 7 show that there was very little difference between the estimates of examiners' adjusted means and regression coefficients for the two cases This means that the regression coefficients for model 8, assuming uniform error, could be used to select those examiners exhibiting wayward marking characteristics, after allowing for centre and sex differences If the samples allocated to these examiners were judged to be typical of the population, as measured by the multiple-choice mark, the scripts could be remarked Where the scripts were drawn from a restricted ability range an investigation would be needed to see whether a blanket adjustment could be applied to the marks of the candidates concerned or whether some remarking was necessary Incorporating these corrective features into a post-examination processing system would be difficult, but not impossible

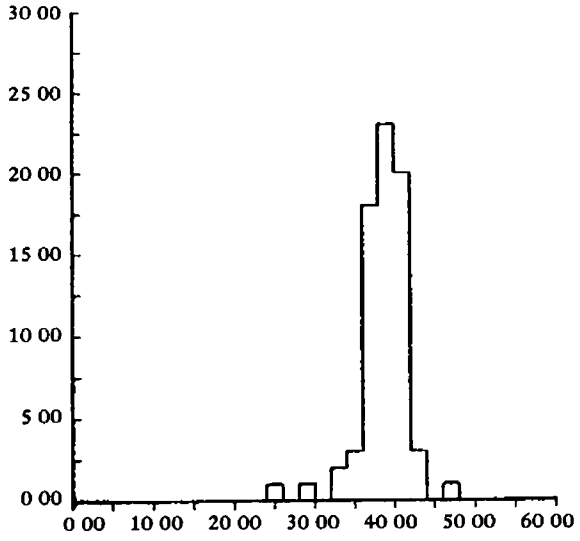

FIGURE 1 DISTRIBUTION OF MCP MEANS FOR MARKER CANDIDATE SAMPLES

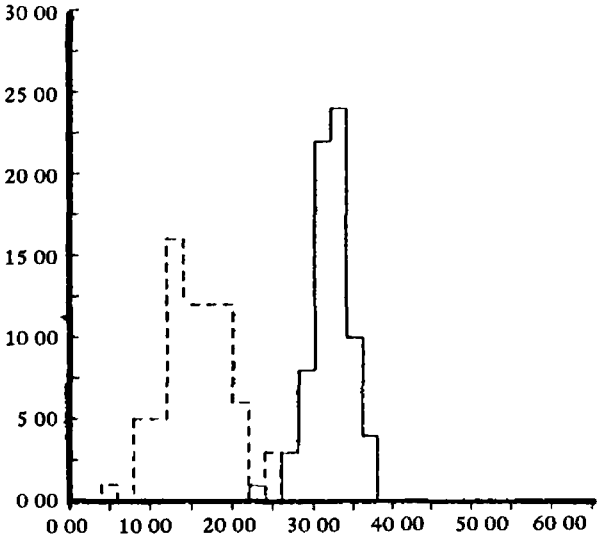

FIGURE 2 DISTRIBUTION OF WP MEANS FOR MARKER CANDIDATE SAMPLES PRE–ADJUSTMENT ( ) AND POST–ADJUSTMENT ( )

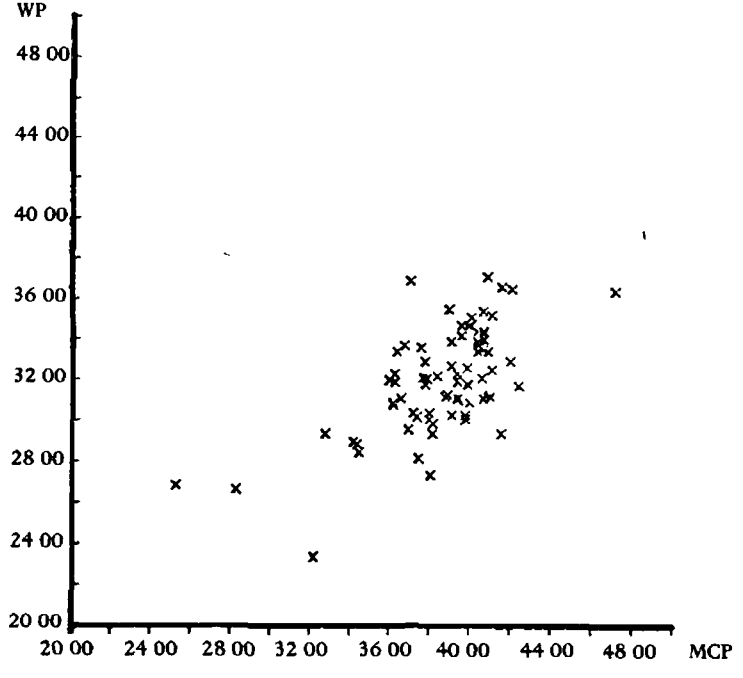

FIGURE 3 SCATTER OF MCP AND WP MEANS FOR MARKER CANDIDATE SAMPLES

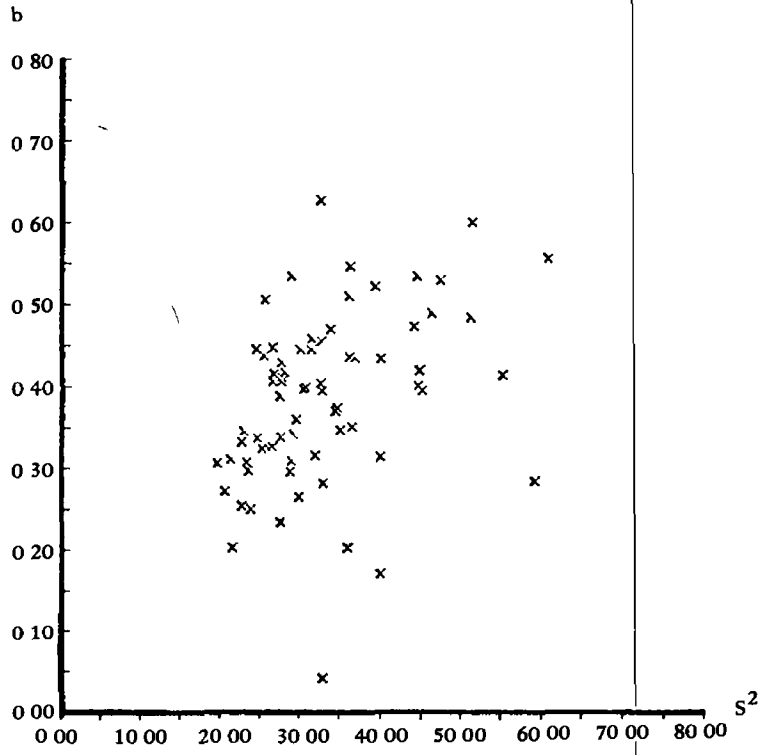

FIGURE 4 MARKER'S ERROR VARIANCE S^2 VERSUS REGRESSION ESTIMATES b

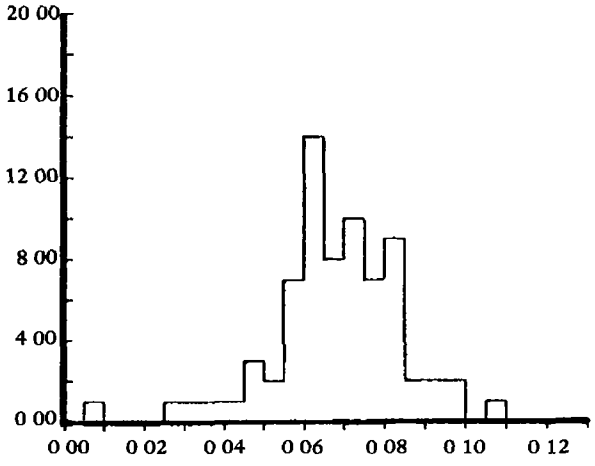

FIGURE 5 DISTRIBUTION OF THE INDEX b_1/v_1

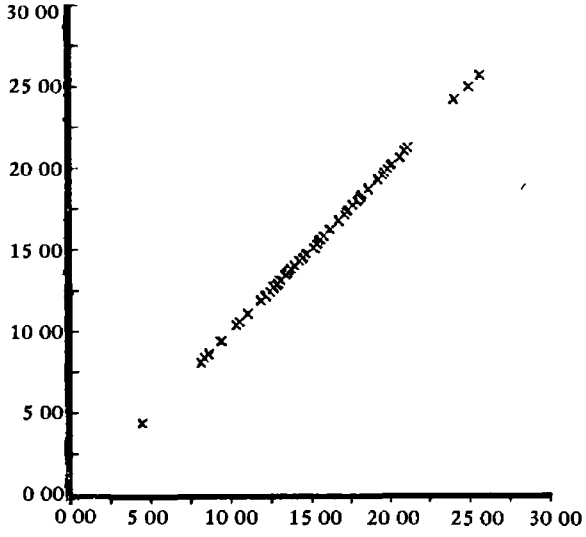

FIGURE 6 ADJUSTED VERSUS UNADJUSTED WP MEANS

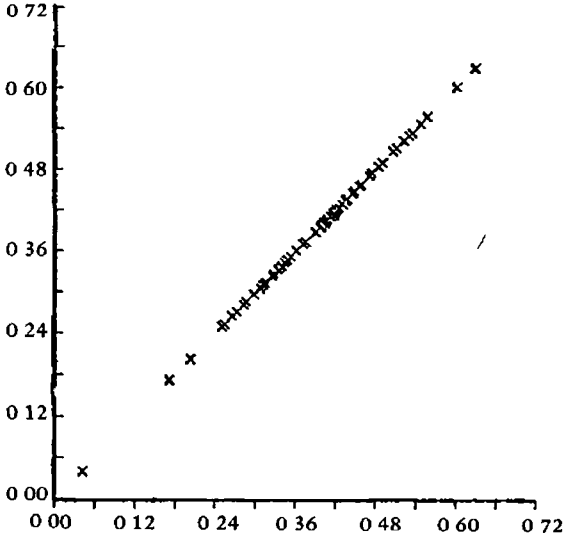

FIGURE 7 ADJUSTED VERSUS UNADJUSTED REGRESSION ESTIMATES

REFERENCES

1 ASHFORD J R and BROWN S Generalised covariance analysis with unequal error variances *Biometrics* 1969 *25* 715 724

2 COCHRAN W G Errors of measurement in statistics *Technometrics* 1968 *10* 637-666

3 COFFMAN W Essay examinations In Thorndike R L (Ed), *Educational Measurement* Washington DC American Council on Education, 1971

4 LINDLEY, D V An experiment in the marking of an examination *Royal Statistical Society Journal Series A* 1961 *124* 285 312

5 NELDER J A and MEAD R A simplex method for function minimisation *Computer Journal* 1965 *7* 308 313

6 O'NEILL R Function minimisation using a simplex procedure Algorithm AS 47, *Applied Statistics* 1971 *20* 338 345

7 OSNES J The influence of extraneous factors on the marking of essays *Scandinavian Journal of Educational Research* 1973, *17*, 161 176