Curriculum, Evaluation and Management Centre

PUBLICATION NO.31

A COMPARISON OF EXAMINATION BOARDS:

A-LEVELS

C. T. FITZ-GIBBON
P. B. TYMMS

Oxford Review of Education
1991

A Comparison of Examination Boards: A Levels

P. B. TYMMS & C. T. FITZ-GIBBON

ABSTRACT *The relative severity of grades awarded by five Exam Boards were compared in 11 subjects at A level using data from the A Level Information System (ALIS) project for 1989. Ordinary least squares regression analysis was used in the first instance to control for prior achievement, gender and ability test scores. Multilevel software was then used to pursue cases where there were apparent differences. There were few instances where discrepancies between Boards were statistically significant having controlled for prior achievement and gender and having employed the appropriate statistical modelling. Further investigation of the ALIS data from 1988 failed to identify consistently severe or lenient Exam Boards over the 2 years except for one Board in one subject.*

INTRODUCTION

Exam Board comparisons have been the subject of much speculation and analysis for more than two decades and are a matter of continuing interest, although the degree of interest has waned somewhat recently in the wake of the current momentous changes in education.

Studies of Exam Board comparability have basically relied on five different techniques (Forrest & Shoesmith, 1985) which go beyond the simplistic comparison of pass rates or proportions awarded certain grades.

The first and potentially most meaningful methodology involves a study of those entrants who took the same exam through different Boards. Problems of non-representative samples and other difficulties were described by Bardell, Forrest & Shoesmith (1978).

A second technique, used by the Boards, involves comparing pass-rates having adjusted for the type of school or college to which the data refer. An example of such a comparison may be found in Kingdon, Wilmut, Davidson and Atkins (1984) in which the authors refer to "the making of many assumptions in order that useful conclusions may be reached". It is also worth noting that the technique involves the use of data aggregated at the level of the institution, a procedure which has been effectively criticised by Aitkin & Longford (1986).

The third technique employs subject pair analysis. The distributions of grades for candidates who take two subjects, say Physics and Chemistry, are compared for the different Boards, the assumption being that the relative distributions should be the same, within sampling variation, across Boards.

A more expensive technique involves the deliberate collection of cognitive data common to all Boards. Ten studies involving such monitor or reference tests were reported in Bardell, Forrest & Shoesmith (1978) but "exercises of this kind have been largely discontinued" (Forrest & Shoesmith, 1985). The practicalities of producing fair and relevant tests were apparently too onerous. The main analysis of the present paper falls into this category of comparison technique.

Cross-moderation is the technique which is now commonly used by the Boards to compare and adjust for differing severities. It is a technique which involves examiners, from different Boards, scrutinising selected scripts, and making considered judgements concerning their relative merits. To quote Bardell, Forrest & Shoesmith (1978) "Cross-moderation methodology is particularly attractive, for it involves the very people who influence the most critical decisions... it has proved surprisingly difficult to design research studies which will result in conclusions of a quantitative kind capable of being translated into action at grading meetings".

A summary of the findings of the many Board comparisons may be found in Appendix B of Forrest & Shoesmith (1985). It is clear from these summaries and from other work (Miles, 1979; Kingdon *et al.*, 1984) that whilst slight severities or leniencies have been noted on occasions the overall picture is one in which there is a broad measure of equivalence between Boards. Jones states in his foreword to Forrest & Shoesmith (1985), which reviews comparability studies since the previous review in 1978;

> ...like the first, it shows that, although variations in standard have been detected and corrected in individual cases, they do not occur as frequently in the context of careful, systematic and collaborative research as they do in the anecdotes of the press and staffroom.

ALIS (A LEVEL INFORMATION SYSTEM)

The data which was used for the investigation in this paper came from the A Level Information System (ALIS) which has been running since 1987 and which developed from the COMBSE (Confidential Measurement Based Self Evaluation) project started in 1983 (Fitz-Gibbon, 1985, 1990, 1991). The system provides feedback concerning A level provisions to school/college departments, Heads/Principals and LEAs. Information is collected from students on the 11 most popular A level courses by way of an extensive questionnaire and ability tests. Once the examination results have been received three sets of reports, in each of the 11 subjects, are sent to the institutions involved. The first reports relate to exam results and provide tables of intake measures, exam grades and exam grades corrected for a variety of intake measures. The second set feed back the reported attitudes of students to their institutions and to their A level subjects as well as dealing with the likelihood of students remaining in education (i.e. going on to Polytechnics and Universities). This measure is also corrected for prior achievement. Finally, the third report deals with classroom processes as reported by students in the questionnaires. The processes are related to the corrected Exam results (ordinary least squares residuals) and to attitudes.

One of the research interests of the project is to try to explain the variation in exam success across departments and to this end data relating to entry by Exam Board are also collected. It is this aspect which is to be followed up in the rest of this paper.

THE POPULARITY OF EXAM BOARDS

Table I and II set out the Boards through which the various institutions entered A

TABLE I. *Exam Boards entered by departments*

	SUBJECT										
School	B	C	E	E	F	G	G	G	H	M	P
or	i	h	c	n	r	e	e	e	i	a	h
College	o	e	o	g	e	n	o	r	s	t	y
1	1	1	1	1	1	1	1	1	1	1	1
2	1	5	1	1	5	1	1	5	1	5	.
3	1	–	–	1	1	1	4	1	1	.	.
4	1	1	1	4	5	1	1	1	1	1	1
5	1	1	1	1	2	1	1	2	1	1	1
6	1	1	1	1	1	1	5	1	1	1	1
7	1	1	1	1	4	1	5	5	1	1	1
9	1	1	4	1	1	1	1	.	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1
11	.	.	.	.	.	.	.	.	.	.	.
12	1	1	4	1	1	1	1	.	1	4	1
13	.	.	4	4	4	.	4	.	4	4	1
14	1	2	.	1	2	2	2	.	2	2	2
15	.	.	.	.	.	.	.	.	.	.	.
16	3	3	3	3	3	.	3	3	3	4	3
17	1	1	1	4	1	.	1	1	1	1	1
18	2	2	5	4	4	2	2	2	2	2	2
19	.	.	1	.	.	.	.	.	.	.	.
20	1	1	.	4	.	.	1	.	1	1	1
21	.	.	.	.	.	.	.	.	.	.	.
22	3	3	.	3	3	.	3	3	3	3	3
23	1	1	1	1	1	1	1	1	1	4	1
24	3	3	1	4	3	1	3	1	3	3	3
25	.	1	.	.	.	.	.	.	1	1	4
26	2	2	2	4	1	.	2	1	1	2	1
27	.	.	.	.	.	.	.	.	.	.	.
28	.	.	.	.	.	.	.	.	.	.	.
29	.	.	.	.	.	.	.	.	.	.	.
30	2	2	1	2	4	.	2	4	1	2	2
31	4	4	.	4	5	4	5	5	4	4	4
32	2	1	4	2	2	.	2	2	1	2	1
33	1	1	.	4	1	.	1	1	1	1	1
34	.	.	.	.	.	.	.	.	.	.	.
35	4	4	4	4	.	4	4	.	.	1	4
36	1	1	3	1	3	1	3	3	3	1	1
37	.	.	.	.	.	.	.	.	.	.	.
38	1	.	3	.	.	.	3	.	1	.	5
39	5	.	.	.	.	.	.	.	.	.	.
40	.	.	.	.	.	.	.	.	.	.	.
41	.	5	.	4	.	.	.	.	.	5	5
42	.	2	.	.	.	.	.	.	4	.	4
43	.	.	.	.	.	.	.	.	.	.	.
44	2	3	.	1	2	.	2	.	1	2	2
45	1	.	4	1	4	1	1	5	1	1	1
46	4	4	4	4	.	1	1	.	1	4	1
47	.	.	.	.	.	.	.	.	.	.	.
48	4	.	4	1	4	1	2	.	1	4	4
49	2	5	3	1	1	1	4	4	4	3	1
50	.	.	.	.	.	.	.	.	.	.	.

TABLE I. *continued*

	SUBJECT										
School or College	B i o	C h e	E c o	E n g	F r e	G e n	G e o	G e r	H i s	M a t	P h y
51	.	.	.	.	.	.	.	.	.	.	.
52	.	.	.	.	.	.	.	.	.	1	.
53	1	1	.	1	1	1	.	.	1	1	1
54	5	1	1	1	5	.	5	4	5	5	5
55	1	1	.	1	3	4	1	3	3	1	1
56	2	.	.	2	2	.	1	2	2	4	2
57	.	.	.	.	.	.	.	.	.	.	.
58	1	5	.	4	4	.	1	4	4	4	1
59	1	1	2	1	1	1	2	1	2	2	1
60	1	3	1	1	3	1	3	3	1	1	3
61	1	1	3	3	1	3	3	1	3	1	1
62	.	5	.	3	4	.	2	3	3	1	1
63	.	.	.	.	.	.	.	.	.	.	.
64	3	4	3	3	3	4	4	3	3	4	1
65	.	.	.	.	.	.	.	.	.	.	.
66	1	3	.	1	3	1	1	3	1	3	1
67	4	3	1	2	3	.	1	3	3	1	3
68	1	3	1	1	3	1	1	3	1	1	1
69	.	3	.	.	3	.	3	.	1	3	3
70	3	3	3	3	1	.	3	.	5	3	1

level subjects in 1989. The Boards are referred to by numbers 1 to 5 throughout this paper.

Among the six Northern LEAs from which we have data on about 4000 pupils the most popular Board for all subjects was Board 1 with 52% of all departmental entries being through that Board. The least popular was Board 5 which accounted for only 6% of departmental entries. Boards 2, 3, and 4 were entered by 11, 17 and 15% of department entries. For most individual subjects this was the general pattern of exam entries. But in French and German Board 1 was less popular than might have been expected and in General Studies there were very few entries except through Board 1.

TABLE II. *Total number of departments using each Board*

Board	Bio	Che	Eco	Eng	Fre	Gen	Geo	Ger	His	Mat	Phy	Tot
1	27	21	17	24	16	23	20	13	29	23	30	243
2	7	5	2	4	5	2	9	4	4	7	5	54
3	5	10	7	6	11	1	9	10	9	6	6	80
4	5	4	8	13	8	4	5	4	5	10	5	71
5	2	5	1	0	4	0	4	4	2	3	3	28

Reasons for these entry patterns are not speculated on here but in passing it should be noted that the figures in Tables I and II are not as clean as they might seem to be. There were cases where the odd candidate, presumably a resit candidate, entered the subject through a different Board and in some instances a particular subjects syllabus is provided by one Board (e.g. Nuffield) but is entered through other Boards. No recognition is made of this practice in the analysis.

Brand loyalty would seem to be quite an important factor when the entry characteristics of schools are considered, the most common entry pattern being for a school to enter A level through two Boards with some entering through one or three Boards. It was rare for an institution to make use of more than three Boards.

This entry pattern would seem to indicate that there are few occasions when departments actually change the Exam Board through which they enter their A level candidates. This pattern is consistent with impressions derived from interviews with Heads of Department: intertia and tradition characterise most 'decisions' about Exam Boards. Caution may also be a factor in that a change of Board may make excessive demands on staff.

THE RELATIVE SEVERITY OF EXAM BOARD GRADES

The comparison of grades awarded by the various Exam Boards in the data set under consideration will be described in a series of stages progressively focussing on apparent discrepancies using fairer and fairer techniques.

For all the investigations the A level grades were converted to a numerical scale where an A grade is given 5 points, a B is given 4 and so on. The average grades in each of the 11 subjects for each Board is shown in Table III. The Table also gives the numbers of entrants (n) on which the averages were based and the standard error (SE) for each average.

The final column gives the probability that the obtained pattern of average grades would have arisen by chance if there were in fact no difference between grades awarded by the Boards and the differences observed were simply due to sampling variation. In six subjects (Biology, Chemistry, Economics, Geography, History and Physics) there were very significant differences between Boards ($p<0.01$). For example in Physics, Board 1 with 418 entrants gave an average grade of 1.97 but in Board 3 with 61 entrants the mean was 2.98, a difference of more than a grade.

In 3 subjects (General Studies, German and Maths) the differences were worth noting but not very significant ($p<0.08$). Whereas in two subjects (English and French) the differences between Boards were small and not statistically significant.

There did not appear to be any consistent pattern in the discrepancies between Boards from subject to subject. For example the average grade for Board 1 in Chemistry was 2.54, equivalent to a C/D and the highest for any Board in Chemistry but in Maths the average for Board 1 was 1.92, just less than a D, the lowest of any Board in Maths.

But, any comparison between Boards cannot be fair unless like is compared with like and no evidence has been presented so far concerning differences which may exist between the entrants for the various Boards. Extensive background data was available on the students. In particular details were available concerning their prior examination success, socio-economic status (SES) and ability scores on maths and verbal tests.

As would be expected theoretically, the ALIS project has repeatedly found the pupil's average O level score (in all subjects entered at O level) to be the best predictor of A level performance with correlations around 0.6. Typically SES has been a poor predictor of A level performance ($r\approx0.1$) and ability test scores have varied from ability test to ability test and from subject to subject. The ALIS/COMBSE

TABLE III. *Average Grades (1989)*

Subject	Board 1	Board 2	Board 3	Board 4	Board 5	p
Biology	2.41	1.69	2.52	2.90	1.75	0.004
SE	0.10	0.22	0.27	0.26	0.37	
n	374	72	44	31	32	
Chemistry	2.54	1.37	2.33	1.43	2.16	0.000
SE	0.10	0.21	0.17	0.33	0.32	
n	366	67	112	23	38	
Economics	2.25	1.42	1.57	0.61	2.45	0.0000
SE	0.11	0.28	0.24	0.20	0.30	
n	293	31	58	42	29	
English	2.55	2.69	2.41	2.37		0.49
SE	0.09	0.21	0.23	0.12		
n	384	62	64	172		
French	1.86	1.62	1.68	1.23	1.55	0.35
SE	0.21	0.24	0.24	0.22	0.27	
n	100	52	69	62	33	
Gen. St.	1.59	2.27	1.41	1.11		0.07
SE	0.07	0.23	0.28	0.22		
n	892	41	37	37		
Geography	2.43	1.51	2.25	2.41	1.95	0.0003
SE	0.12	0.16	0.21	0.21	0.28	
n	235	109	83	41	42	
German	1.91	1.57	2.94	1.50	1.91	0.048
SE	0.27	0.33	0.30	0.80	0.56	
n	45	28	35	8	11	
History	1.70	2.16	2.52	2.06	1.67	0.0016
SE	0.10	0.17	0.16	0.30	0.29	
n	371	90	77	31	39	
Maths	1.92	2.31	2.42	2.37	2.15	0.030
SE	0.08	0.15	0.22	0.20	0.20	
n	614	170	69	107	79	
Physics	1.97	1.59	2.98	1.76	2.23	0.0006
SE	0.10	0.21	0.21	0.32	0.31	
n	418	73	61	34	40	

project has tried a variety of high level ability tests throughout its development. The correlations have been as low as 0.21 (n=122) between Raven's Advanced Progressive Matrices with A level English Literature and as high as 0.58 (n=514) between Physics and the International Test of Developed Abilities Maths (ITDAM). To date the ITDAM has proved to be a satisfactory control for A level Maths and Science but an equivalent test for other A level subjects has yet to be found although a vocabulary test developed by Fitz-Gibbon did correlate 0.47 (n=205) with A level English Literature in 1989.

Table IV shows the exam grades corrected for prior achievement (Mean O level/GCSE grade and gender) and Table V shows the same data corrected for the two ability test scores (maths and verbal) as well as SES and gender. For both tables ordinary least squares regression (OLS) analysis was used and in both cases the variables were added 'stepwise' into the equation with a cutoff point of p=0.05. Consequently not all the explanatory variables were employed for all the regressions.

TABLE IV. *OLS Residuals controlling for GCE/GCSE & Gender (1989)*

Subject	Board 1	Board 2	Board 3	Board 4	Board 5	*p*	R
Biology	0.11	−0.23	−0.11	0.82	−0.25	0.0065	0.65
SE	0.07	0.21	0.20	0.21	0.31		
n	374	72	44	31	31		
Chemistry	0.22	−0.37	−0.07	−0.16	−0.55	0.0006	0.67
SE	0.11	0.18	0.12	0.30	0.20		
n	364	65	112	23	38		
Economics	0.35	−0.87	−0.44	−1.27	0.18	0.0000	0.55
SE	0.09	0.27	0.19	0.14	0.32		
n	292	31	58	42	28		
English	0.14	0.18	−0.18	0.19		0.35	0.52
SE	0.08	0.21	0.17	0.10			
n	384	46	64	170			
French	0.13	−0.06	−0.18	−0.60	0.12	0.035	0.60
SE	0.17	0.16	0.17	0.18	0.23		
n	99	52	69	62	33		
Gen. St.	0.03	0.75	−0.30	−0.07		0.02	0.56
SE	0.05	0.21	0.26	0.20			
n	890	40	37	36			
Geography	0.34	−0.72	−0.03	0	−0.08	0.0000	0.55
SE	0.10	0.14	0.17	0.19	0.23		
n	234	104	83	41	39		
German	−0.21	−0.51	0.39	1.64	0.43	0.028	0.61
SE	0.22	0.31	0.22	0.41	0.36		
n	44	28	35	4	11		
History	−0.21	0.26	0.48	0.23	0.07	0.002	0.48
SE	0.09	0.15	0.14	0.30	0.24		
n	349	90	77	30	38		
Maths	−0.03	0.25	0.44	0.16	0.27	0.064	0.55
SE	0.07	0.13	0.17	0.18	0.18		
n	613	159	69	106	79		
Physics	−0.06	−0.18	0.60	0.13	0.51	0.0016	0.69
SE	0.07	0.18	0.16	0.30	0.23		
n	417	65	61	34	40		

Tables IV and V represent corrections on two quite different bases. In some respects Table IV presents the fairer corrections because it is based on greater numbers and because it takes account of more pupil level variance than Table V, except in the case of A level Physics. On the other hand, the prior achievement scores are themselves based on grades awarded by the Exam Boards. And if there are differences between Boards then to control for the Boards own awards would seem to be a suspect procedure.

Both Tables indicate that there were differences between the Boards in Chemistry, Economics, Geography, History and Physics. For Biology the differences in Table IV were significantly different but not in Table V. This may be because there were fewer cases on which to base Table V than IV and also because the controls were less effective in Table V. English also presented different pictures in the two Tables. In Table IV, with the best controls and highest number of entrants, there was no

TABLE V. *OLS Residuals controlling for Ability, SES & Gender (1989)*

Subject	Board 1	Board 2	Board 3	Board 4	Board 5	*p*	R
Biology	0.02	−0.40	0.29	0.86	−0.49	0.10	0.38
SE	0.11	0.26	0.39	0.52	0.39		
n	209	52	15	10	24		
Chemistry	0.23	−0.58	−0.63	−0.28	−0.69	0.0012	0.48
SE	0.11	0.21	0.28	0.87	0.28		
n	234	56	28	5	18		
Economics	0.36	−1.49	−1.02	−1.07	0.37	0.0000	0.32
SE	0.13	0.29	0.28	0.18	0.33		
n	181	6	11	33	25		
English	0	−0.14	0.21	0.56		0.02	0.37
SE	0.11	0.47	0.50	0.11			
n	240	10	11	98			
French	0.23	−0.32	−0.20	−0.53	0.12	0.27	0.38
SE	0.25	0.25	0.44	0.26	0.36		
n	61	40	17	18	18		
Gen. St.	−0.04	0.75		−0.69		0.0066	0.53
SE	0.07	1.15		0.54			
n	578	20		6			
Geography	0.35	−0.66	0.07	0.64	0.24	0.0003	0.33
SE	0.15	0.17	0.28	0.34	0.31		
n	128	74	38	15	26		
German	−0.11	−0.40	0.76	1.38	−0.14	0.49	0.43
SE	0.33	0.38	0.87	0.55	0.41		
n	27	21	5	3	4		
History	−0.17	0.41	0.55	0.06	−0.48	0.04	0.26
SE	0.11	0.19	0.29	1.02	0.39		
n	260	59	22	5	20		
Maths	0.11	−0.17	0.34	0.58	−0.04	0.13	0.46
SE	0.08	0.19	0.35	0.27	0.18		
n	364	100	16	52	62		
Physics	−0.15	0.05	0.67	0.41	−0.57	0.0016	0.69
SE	0.11	0.27	0.27	0.39	0.33		
n	210	45	22	17	19		

significant difference between Boards but in Table V Board 4 appeared to be leniently graded. General Studies was predominantly entered through Board 1 and the data pointed to Board differences in Tables IV and V. However the significance level was only 2% in Table IV. In Table V the difference was mainly due to an apparent leniency in Board 2 which was only taken in two schools and it would be unreasonable, therefore, to conclude that Board 2 was lenient in its grading rather than that the two schools had been successful in their teaching. This is of course an instance of the general problem of a confound of school effects with Board effects. There was some evidence for Board differences for German in Table IV but not Table V. In neither of the Tables was there a statistically significant difference between Boards for maths.

Before moving to a more detailed look at those cases where there was clear evidence for differences it should be noted that Tables IV and V do not give any support for the idea that there is an 'easy' or a 'hard' Board in any general sense. For instance going down Table IV and picking out the most lenient Boards for each subject gives the list 4, 1, 1, 4, 1, 2, 1, 4, 3, 3, 3. Whilst the same exercise picking out the most severely graded Boards produces the list 5, 5, 4, 3, 4, 3, 2, 2, 1, 1, 2.

If there is no such thing as an 'easy' Board or a 'hard' Board overall but there are differences within individual subjects and if this is true for 16+ exams as well as for A level then it makes sense to use GCE and GCSE grades as controls when looking at Board differences at A level provided average grades are used as controls. The point is that when averages are used the differences which do exist will be largely ironed out. The possibility that schools quite often chose to enter O level/GCSE subjects through a variety of Boards in 1987 adds to the plausibility of the argument that averaging the grades out will tend to give a consistent measure of prior achievement.

MULTI LEVEL ANALYSIS

Those subjects for which marked differences were found between Boards were examined in more detail using the ML3 software of Prosser, Rasbash & Goldstein (1989). This software allows separate regression equations to be formed for each school/college and the differences in the equations can be explained by measurable school/college level variables. This is exactly the kind of model required for a comparison of exam Boards since the Board entered can be modelled as a school level variable. Furthermore, in so doing some of the problems associated with OLS regression are overcome (Aitkin & Longford, 1986; Goldstein, 1987; Cuttance, 1985, Raudenbush & Bryk, 1986) providing more accurate estimates of Exam Board differences and their errors of measurement.

The exact procedure followed for the multilevel analysis was as follows: First the variances at pupil level and at school level were estimated with no explanatory variables in the model. Second the pupil variation was modelled using the average GCE/GCSE grade (AVOG) and gender. The intercept and the AVOG coefficient (slope) were allowed to vary between schools. In no case was there evidence of AVOG coefficient variation and it was fixed for all future models. The intercept did vary across schools. Third the average AVOG for each school was entered as a school/college level variable. That is to say the model was tested for a contextual effect. This has often been found to be an important variable in school effectiveness studies (e.g. Willms, 1986, 1988) but it was not important this investigation and it was dropped from the models. Fitz-Gibbon (1991) found a similar lack of contextual effect in her multi-level analysis of the 1988 ALIS data. The absence of a contextual effect could be because there were rather small numbers of A level candidates in some schools or perhaps because there was in fact no contextual effect present at A level. More extensive data sets should help to throw light on this interesting question. The fourth stage in modelling was to introduce school level dummy variables to model the Exam Boards. The results of this analysis appear in Table VI. Finally the software was used to obtain estimates of the likelihood of various contrasts appearing by chance. Details from this analysis appear in Table VII.

Table VI does not give details of Board 1 since all other Boards were being compared to it as a standard. In other words Board 1 may be regarded as having a score of 0. The Table gives a very similar patern of results to Table IV but the advanced software gives more efficient estimates of Board differences and more realistic but higher estimates for the errors. The results in Table VI are reorganised into Table VII so that comparisons between Boards may be made more easily.

TABLE VI. *Dummy Coefficients from ML3 (GCE/GCSE & Gender) (1989)*

Subject	Board 2	Board 3	Board 4	Board 5
Biology	−0.41	−0.13	0.85	−0.31
SE	0.28	0.29	0.32	0.37
Chemistry	−0.40	−0.21	−0.33	−0.77
SE	0.33	0.22	0.37	0.34
Economics	−1.43	−0.61	−1.43	0
SE	0.55	0.36	0.37	0.70
Geography	−0.93	−0.39	−0.27	−0.51
SE	0.30	0.30	0.40	0.40
History	0.28	0.51	0.27	0.21
SE	0.39	0.30	0.43	0.51
Physics	−0.41	0.72	0.25	0.48
SE	0.33	0.28	0.35	0.37

Biology. It would appear that Board 4 was graded leniently in 1989 to the tune of between 0.85 grades (against Board 1) and 1.26 grades (against Board 2). It would be unfair to make an definitive statement about Board 4's Biology unless there were evidence of the pattern being repeated in other years and in other areas but the evidence for 1989 in the data set under investigation is fairly convincing. Although it is not possible with the data available to distinguish between the confounding of a Board effect and a school effect, in this case five schools took Biology through Board 4 and the differences were large even for individual school effects and so it would seem more likely to be a Board effect than otherwise.

Chemistry. There was some evidence that Board 5 was more severely graded than Board 1 (0.77 of a grade), but the significance level of the finding was low and when there are so many possible comparisons to be made between Boards in 11 different subjects it is as well to be conservative when identifying differences and regard this particular difference as something to be recalled if further evidence for the same Board differences appear in other years. Nevertheless, five different schools did take the Board 5 exam.

Economics. Boards 5 and 1 were equivalent in their severity as were Boards 2 and 4 but the latter pair were more harsh by 1.43 grades than the former pair. The differences were significant only for comparisons between the most popular Board, 1, and Boards 2 and 4.

Geography. As with Chemistry the evidence for differences was slight with Board 2 apparently being more harshly graded than Board 1 by 0.93 of a grade. With nine schools making use of Board 2 it would seem unlikely that the difference was due to a school effect but it will be important to look for stability over the years.

History. There were no significant differences between Boards in History.

TABLE VII. *Differences between Boards (1989)*

Biology

	Board	2	5	3	1
−0.41	2				
−0.31	5				
−0.13	3				
0	1				
0.85	4	*	*	*	*

Chemistry

	Board	5	2	4	3
−0.77	5				
−0.40	2				
−0.33	4				
−0.21	3				
0	1	*			

Economics

	Board	2	4	3	5
−1.43	2				
−1.43	4				
−0.61	3				
0	5				
0	1	*	*		

Geography

	Board	2	5	3	4
−0.93	2				
−0.51	5				
−0.39	3				
−0.27	4				
0	1	*			

History

	Board	1	5	4	2
0	1				
0.21	5				
0.27	4				
0.28	2				
0.51	3				

Physics

	Board	2	1	4	5
−0.41	2				
0	1				
0.25	4				
0.48	5				
0.72	3	*	*		

$^{*}p<0.05$.

Physics. Board 3 appeared to be more leniently graded than Boards 2 and 1 by 1.13 and 0.72 grades respectively.

EVIDENCE FROM 1988

As has already been noted any inter-Board comparisons should really be made on a longitudinal basis and clear statements about relative severity and leniency can only be made following such an analysis. Such data will gradually accumulate as the ALIS project continues. Meanwhile it is possible to look at the 1988 data for evidence of stability of the Board differences found in the 1989 data. (The COMBSE project started in 1983 but only for English and maths and only for a few institutions.)

Table VIII presents the OLS analysis of the 1988 exam results for those subjects for which differences were found in the 1989 data. The format of the Table is the same as for Table IV except that the number of institutions (N) is included.

The sample was smaller than in 1989 and was particularly thin for some Boards but the overall pictue was as follows. There was no evidence of statistically significant Board differences in Chemistry and Economics and those differences which did exist did not match the 1989 data. Board 5 in Chemistry was no longer the most severe and Board 1 in Economics was no longer the most lenient.

There were significant differences, at the 0.04 and 0.01 levels, between Boards for Biology and for Physics but again the pattern of differences did not match the 1989 pattern. In Biology Board 3 was lenient but in 1989 it was Board 4 which was lenient. However, data was only available for 13 students in 3 institutions for Board 4 in 1988. In Physics there is evidence for some leniency on the part of Board 3 as in 1989 but only for 26 entrants in 3 institutions whilst Board 2 with 83 candidates has reversed its position from most severe in 1989 to most lenient in 1988.

Only for Geography did there appear to be consistent Board differences over two years. Board 2 appeared to be harsh in 1989 and the same pattern showed up in 1988 using OLS analysis. The data for Geography in 1988 were analysed via ML3 using the same process described earlier in this paper. The results of that analysis are presented in Table IX which confirms the relative severity of Board 2 for the two years amounting to between 0.54 and 1.20 of a grade more than Boards 1, 3 and 4. But once again a note of caution should be sounded. To have found a single Board with extreme grades, even after correction, out of five Boards in 11 subjects is hardly surprising although consistency over 2 years does make the finding more important.

DISCUSSION

The above investigation sought to locate differences between Boards. It is important to question the assumptions on which the investigation was based and to look at implications which the results might have.

The most important underlying assumption is that it is possible to 'correct' exam results using prior achievement measures and then to make fair comparisons. In general terms this assumption must be treated with caution. There are problems in regression analysis in treating exam grades as a continuous scale and there is the difficulty of under correction to bear in mind. Nevertheless prior achievement (GCE/GCSE) has proved to be a good control variable and in view of the restricted

TABLE VIII. *OLS Residuals controlling for GCE/GCSE & Gender 1988 data*

Subject	Board 1	Board 2	Board 3	Board 4	Board 5	*p*	R
Biology	−0.05	0.01	0.85	0.15	−1.73	0.04	0.61
SE	0.09	0.20	0.41	0.35	0.65		
n	241	73	18	13	3		
N	29	6	3	3	1		
Chemistry	0.09	−0.19	−0.36	−0.28	−0.26	0.12	0.65
SE	0.07	0.16	0.18	0.47	0.43		
n	454	84	57	2	20		
N	23	5	5	1	2		
Economics	−0.04	−0.38	−0.03	0.26	0.39	0.18	0.53
SE	0.09	0.22	0.29	0.21	0.29		
n	321	38	21	43	21		
N	16	4	4	5	3		
Geography	0.13	−0.45	0.30	0.31	−0.45	0.005	0.49
SE	0.10	0.16	0.22	0.40	0.21		
n	253	76	61	20	61		
N	19	6	7	2	6		
Physics	−0.10	0.45	0.44	−0.02	−0.08	0.01	0.64
SE	0.07	0.14	0.31	0.39	0.23		
n	447	83	26	15	51		
N	27	6	3	2	3		

n number of students.
N number of institutions.

TABLE IX. *Differences between Boards (1988)*

Geography

	Board	2	5	1	3
−0.77	2				
−0.27	5				
0	1				
0.16	3	*			
0.43	4	*			

*$p<0.05$.

ability range of A level students it is a very good control variable indeed and far better than most variables used in recent school effectiveness studies (e.g. Rutter *et al.*, 1979; Mortimore *et al.*, 1988; Willms, 1986; Raffe, 1988). The problem of exam grades not representing a continuous scale can be dealt with to some extent by converting grades to more precise z scores (corresponding the Centre of Gravity of each grade) but this appears to make little difference to A level school effectiveness scores (Tymms, 1986). But a more serious difficulty with the investigation is that Board differences are confounded with school/college differences. This is a particular problem if there are few schools taking an individual Board for which discrepancies are found and indeed it is the case that in the data set investigated here differences were found for minority Boards. The situation would be greatly helped by investigation of other data sets.

Undercorrection is sure to be present in this kind of analysis and it is well worth noting that in 1989 in Biology Board 4 had the highest grades prior to adjustment as well as after adjustment. In Economics Boards 5 and 1 similarly had the highest grades before and after and in Geography Board 2 had the lowest grades before and after. In Physics Board 3 had the highest grades before and after adjustment.

It is possible that the differences which have been identified were simply cases where extreme groups in terms of output measures have been spotted and the correction procedures have not been sufficient to be fair to these outliers. However, as has already been stated the control variables were good and if there had been serious undercorrection one might have expected a contextual effect to have appeared.

But behind all these problems is the difficulty of random assignment which is implicitly assumed in all the analyses. This is an insurmountable problem in this kind of investigation and it is finally necessary to weight up the situation and decide on balance if the problems with the analysis are sufficient to dismiss the findings or vice versa.

In the series of analyses presented above the exam results in 1989 from five different Boards in 11 different A level subjects have been compared. Within each subject this represents ten possible comparisons; a total of 110 comparisons. Marked differences (more than 1 grade with $p<0.05$) were found in only three subjects. In Biology Board 4 appeared more leniently graded than all other Boards whilst in Economics Board 1 was more lenient than Boards 2 and 4 and in Physics Board 3 was more lenient than Boards 2 and 1. Slight differences were found in Chemistry with Board 5 being severely graded and in Geography with Board 2 being severely graded. The more limited 1988 data showed no evidence of consistent inter-Board discrepancies except in Geography where Board 2 was severely graded over the 2 years.

These differences are not trivial but seen in the context of all possible differences they amount to small variations and it is a credit to the Boards that they have been able to maintain such consistent grading schemes.

The few marked differences which have been identified do appear to be real differences but it is only by looking at more data sets from different years and from other geographical areas that any firm view can be established. The findings are clearly of interest both to schools/colleges and to the Boards and one might expect that information of this sort could alter policy decisions. At this stage the data together with Board identities will only be made available to all those in the ALIS project. In the meanwhile it is to be expected that the Boards themselves will continue to review their own grading procedures and possible alterations in grading may follow. Similary institutions within ALIS may decide to review their entry policies. It is hoped to feed back similar information to institutions within ALIS on a regular basis so that their decisions may be taken on an informed basis.

But should all Boards or all syllabuses within Boards award grades which correspond to the input characteristics of their entrants? Surely not. It may be that a certain syllabus encourages higher standards of work or attracts more dynamic departments through its stimulating approach, in which case a difference in grading, calculated on the basis of input characteristics, would be expected and justifiable. But if this really is the case then schools and colleges presumably should know it and institutions of Higher Education should be aware of these differences so that they may be informed and able to make fair decisions with regard to entry policy.

Once one begins to consider questions of differing contents and aims between syllabuses the situation becomes almost irresolvable, and certainly moves into the realm of value decisions which are beyond statistical treatment.

CONCLUSION

The A level examining system acts as gatekeeper to Higher Education and the professions and as such it is important that it is, and is seen to be, a fair system. We have worked here on only one aspect of fairness, the possibility that the various Boards award grades with differing degrees of severity. It is not uncommon to hear rumours amongst teachers about one Board's exam being particularly hard or easy. Such rumours can be divisive and without solid data they are difficult to deal with. However, the evidence presented in this paper demonstrates that in the vast majority of cases there is an impressive degree of consistency amongst the Boards. A few cases are identified where there appears to have been discrepancies but evidence for consistent differences over 2 years was lacking except in the case of one Board in one subject. This standardisation must, we presume, be due to strenuous efforts by the Examining Boards, efforts which seem to have been commendably effective.

There are other issues of fairness, for example the possibility of unequal difficulties across subjects and failures to certify genuine achievements (as evidenced by the 30% failure rate), which are not dealt with here, but which should concern everyone involved in educational policy.

REFERENCES

AITKIN, M. & LONGFORD, N. (1986) Statistical modelling issues in school effectiveness studies, *Journal of the Royal Statistical Society, Series A*, 149(1), pp. 1–43.

BARDELL, G.S., FORREST, G.M. & SHOESMITH, D.J. (1978) *Comparability in GCE: a review of the boards' studies 1964–1977* (Manchester, JMB on behalf of the GCE examining boards).

CUTTANCE, P. (1985) Methodological issues in the statistical analysis of data on the effectiveness of schools, *British Educational Research Journal*, 11(2), pp. 163–179.

FITZ-GIBBON, C.T. (1985) A-level results in comprehensive schools: the COMBSE project, year 1, *Oxford Review of Education*, 11(1), pp. 43–58.

FITZ-GIBBON, C.T. (1990) An Up-and-Running Indicator System, in: C. T. FITZ-GIBBON (Ed.) *Performance Indicators: a BERA Dialogue* (Clevedon, Avon, Multilingual Matters).

FITZ-GIBBON, C.T. (1991) Multilevel modelling in an indicator system, in: RAUDENBUSCH, R.W. & WILLMS, J.D. *Schools, Pupils and Classrooms: International Studies of Schooling from a Multilevel Perspective* (London & New York, Academic Press).

FORREST, G.M. & SHOESMITH, D.J. (1985) *A Second Review of GCE Comparability Studies* (Manchester, Joint Matriculation Board).

GOLDSTEIN, H. (1987) *Multilevel Models in Educational and Social Research* (London, Griffin).

KINGDON *et al.* (1984) *Report of the Inter-Board Comparability Study of Grading Standards in Advanced Level* (London, University of London Schools Examination Board).

MILES, H.B. (1979) *Some factors affecting attainment at 18+* (Oxford, Pergamon Press).

MORTIMORE, P. *et al.* (1988) *School Matters* (Wells Somerset, Open Books).

PROSSER, P., RASBASH, J. & GOLDSTEIN, H. (1990) *ML3 Software for Three-level Analysis* (Institute of Education, University of London).

RAFFE, D. (1988) Making the Gift Horse Jump the Hurdles, *British Journal of Education and Work*, Vol. 2, No. 3.

RAUDENBUSH, S. & BRYK, A.S. (1986) A hierarchical model for studying school effects, *Sociology of Education*, 59, pp. 1–17.

RUTTER, M. *et al.* (1979) *Fifteen Thousand hours* (Somerset, Open Books).

TYMMS, P.B. (1986) 'An examination of the A level Results in one area', *Unpublished submission in part fulfillment of requirement for degree of MEd* (University of Newcastle upon Tyne).

WILLMS, J.D. (1988) A Longitudinal Hierarchical Linear Model for Estimating School Effects and their Stability, *Journal of Educational Measurement* (in press).

WILLMS, J.D. (1986) Socal Class segregation and its relationship to students' examination results in Scotland, *American Sociological Review*, 6(3), pp. 289–306.

Correspondence: P. B. Tymms, Moray House College, Holyrood Campus, Holyrood Road, Edinburgh EH8 8AQ, United Kingdom.