J. R. Statist. Soc. A (1996)
159, *Part* 1, *pp.* 149–163

Using Examination Results as Indicators of School and College Performance

By HARVEY GOLDSTEIN† and SALLY THOMAS

Institute of Education, London, UK

[Received August 1994. Revised January 1995]

SUMMARY

A current requirement for secondary schools in England and Wales, associated with the so-called '*Parents' Charter*', is that each school is required to publish its average General Certificate of Secondary Education examination results. Every institution with A-level General Certificate of Education candidates is also required to do this. Additionally, the government arranges every autumn for a national 'league table' of all these results to be published in the national press. A principal official justification for this policy is that it will help parents to choose schools for their children on the basis of how well the school is seen to be performing. The paper argues that institutional comparisons based on average, unadjusted, examination results are inadequate and potentially misleading for several reasons. Aggregate data obscure important information; the failure to take account of prior achievement leads to inaccurate and misleading inferences about school differences and they are always out of date because they refer to a cohort who began attending the institutions several years earlier. An alternative 'value-added' analysis of A-level results will be presented based on individual student data and adjusted for intake achievement. The results illustrate the inadequacies of the current procedure but they also demonstrate that *any* attempts to use examination results to judge the comparative 'effectiveness' of schools and other educational institutions have inherent problems which severely limit the usefulness of such a system for accountability. These results suggest the possibility that better uses for value-added comparisons are as screening instruments to identify institutions for further investigation.

Keywords: EXAMINATION RESULTS; INSTITUTIONAL COMPARISONS; MULTILEVEL MODELLING; PERFORMANCE INDICATORS; SCHOOL EFFECTIVENESS; VALUE ADDED

1. INTRODUCTION

In several areas of public service, in the UK and elsewhere, there is considerable interest in constructing indicators to measure the performance of those services. In health, for example, there are attempts to measure such things as in-patient waiting times and to compare these between institutions. In education the *Parents' Charter* (Department of Education and Science, 1991) requires the publication of examination and national curriculum test results. This is part of a general initiative by the Conservative government elected in 1987 and re-elected in 1992 to promote the use of indicators by which institutions can be compared and hence their performance evaluated. It requires, among other things, that comparative 'league tables' of examination and national curriculum test results be published for every educational

†*Address for correspondence*: Institute of Education, University of London, 20 Bedford Way, London, WC1H 0AL, UK.
E-mail: hgoldstn@ioe.ac.uk

 0035–9238/96/159149

institution and authority. The requirement to do this for key stage 1 (7-year-old) and key stage 3 (14-year-old) students has since been dropped but since the autumn of 1992 the results of General Certificate of Secondary Education (GCSE) examinations taken at the end of compulsory schooling and the results of A-level General Certificate of Education (GCE) examinations have been published in this form, with comparative rankings appearing in the national and local press. The stated intention in the *Parents' Charter* is that these tables should be used by parents and others in choosing schools and colleges. It is the purpose of the present paper to investigate the properties of such tables and in particular to compare them with analyses of examination results which attempt to compare institutions after adjusting for the intake achievements of the students that they receive.

To evaluate the usefulness and fairness of these league tables the present authors, together with the late professor Desmond Nuttall, collaborated with *The Guardian* newspaper to carry out a survey of institutions to obtain examination results which could be analysed by using different methods. Institutions with students following courses leading to advanced and advanced supplementary level (A- and AS-level) GCE examinations were chosen because it was possible also to obtain the results of GCSE examinations taken generally 2 years previously. The A-level examinations are taken for the most part by students in year 13, i.e. the school year when they reach the age of 18 years, and generally in up to four subjects. The first survey was carried out in 1992 and published in October of that year (Guardian, 1992). This used only average results at A-level and at GCSE. In 1993 (Guardian, 1993) a second survey of schools was published based on the analysis of individual student A-level and GCSE results. It is this latter analysis which is described in this paper.

2. DATA COLLECTION

Institutions with A-level candidates were approached in March 1993 to ask for their participation in a survey to select the summer 1993 A-level and AS-level results for each of their students together with the GCSE results for the same set of students. The number of institutions which agreed to participate was 436 which is estimated to be 15% of the total possible. A breakdown by type and status is given in Table 1.

Further data were collected on the number of GCSE examinations taken, the gender of the student, the gender composition of the institution and the age of the student. Separate results for 'general studies' examinations were also obtained but these are not used in the present analysis.

TABLE 1
Numbers of institutions by type and status

Institution type	*Number*	*Institution status*	*Number*
Further education	24	County	254
Tertiary	14	Voluntary	50
Sixth-form college	7	Grant maintained	72
Comprehensive	262	City technical college	0
Selective	99	Private	53
Other	27	Other	4

TABLE 2
Conversion of grades to scores for GCSE and A–AS-level examination subjects

GCSE results		*A–AS-level results*	
Grade	*Score*	*Grade*	*Score*
A	7	A	10 (5)
B	6	B	8 (4)
C	5	C	6 (3)
D	4	D	4 (2)
E	3	E	2 (1)
F	2		
G	1		

The institutions completed precoded forms and were asked to return these to *The Guardian* by the end of September 1993. They were sent to a data processing bureau for transfer to disc and were also subjected to careful editing with queries referred back to institutions. The data set used in the present analysis includes only records with complete information for students in the 17–19 years age range on September 1st, 1992, 1.6%, 91.2% and 7.2% of the total at each age. In addition, institutions with fewer than five candidates were excluded. Out of the original participants, the data for 425 were analysed for *The Guardian*'s report. Only 325 out of these, with 21 654 students, had all the institution level information; it is principally the 1992 A–AS examination score which is missing since this had to be obtained from the data supplied to the media by the Department for Education in 1992 and these data were incomplete.

The scores for GCSE and A–AS-level were constructed by transforming the grades as in Table 2.

For the GCSE results the total score for the eight best grades was chosen and for A–AS-level the total scores were computed for each student, not counting examinations that were retaken. The choice of the eight best GCSE results was dictated by an attempt to deal with varying entry policies for institutions. We discuss the limitations of this choice later.

Table 3 shows a comparison of the subsample used for the present analysis with the full *Guardian* data set and a comparison of the latter with the national results for 1992. The participating institutions appear to have a slightly lower average score

TABLE 3
Mean score for the 1993 subsample of participating institutions and comparisons of this subsample with the omitted institutions and national data†

1993 subsample	15.14
1993 subsample—omitted cases	−0.27 (0.43)
1992 subsample—national average	−0.99 (0.23)

†Standard errors are given in parentheses.

than the national average but there is little difference between the subsample and the full data set.

3. MODELS FOR COMPARING INSTITUTIONS

Educational institutions in England and Wales, as in most other countries, vary in terms of the educational achievements of their students. This results from many factors. Some institutions are formally selective and therefore admit students according to their existing achievements or performances, whereas others may be selective *de facto* for example by virtue of being fee paying or enjoying a high reputation which ensures that they can choose between applicants. It is widely recognized that, because earlier achievement is such a powerful predictor of subsequent achievement, comparisons of 'output' achievements among institutions should take account of intake variation if the contribution to achievement of the institution is of interest. In other words, it is the *progress* made by students from the time that they enter an institution to the time that they leave that should form the basis for comparing the institutions.

There have been several studies, often referred to as 'school effectiveness' research, which have demonstrated the importance of this (see for example Scheerens (1992)). In statistical terms the final examination result is the response and the initial achievement measures are explanatory variables or covariates. A simple such model can be written as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}. \tag{1}$$

Equation (1) expresses the A–AS examination score Y for the ith student in the jth institution as a simple linear function of the GCSE score X plus a term or 'effect' u_j for the jth institution together with the student residual e_{ij}. It is possible to regard the u_j as separate parameters to be estimated, but this becomes practically infeasible with a large number of institutions and is also inefficient for those institutions with small numbers of students. Furthermore, in general we wish to study the nature of the between-institution variation and the extent to which it is explained by further factors, such as gender and socioeconomic status, and we may also wish to regard the institutions as a sample from which to make inferences about a larger universe. For these reasons we choose to regard the u_j as random so that equation (1) becomes a 'multilevel' model with a two-level hierarchy of students nested within schools. It can also be viewed from a Bayesian perspective as a linear model with the u_j assumed exchangeable (Lindley and Smith, 1972). Aitkin and Longford (1986) discuss the use of this model for school effectiveness studies and Goldstein (1987, 1995) gives a general account of multilevel models. For providing comparisons between schools we require estimates $\hat{u}_j$ which can be viewed either as Bayesian posterior estimates or as regression predictions of the unknown u_j given the responses and the model parameter estimates.

Raudenbush and Willms (1995) make an important distinction between two principal kinds of comparison of institutions. One is where we wish to compare expected performances between institutions conditionally on certain student characteristics such as their prior achievement scores or social background.

Raudenbush and Willms label them type A comparisons or effects. The *reasons* for any differences may be due, for example, to the student composition or the institution's internal organization, and such comparisons may be of interest to students choosing institutions. A second kind of comparison is where we are interested in conditioning on student characteristics plus any other variables, such as the student composition of an institution or its level of learning resources, so that the effects of internal institutional policies can be isolated. Raudenbush and Willms term these type B analyses or effects. Also of interest is whether there are interactions between any of these variables, e.g. whether certain kinds of students do well in certain kinds of school.

It is the type B comparisons which are of major interest to researchers and this implies the collection and analysis of a wider range of variables than if we wish simply to make predictions for individual students with particular characteristics in different institutions. The present data were collected principally to study the characteristics of type A comparisons, and in particular to examine how far these can be used to guide the choice of institution.

We have carried out a preliminary series of analyses to determine a parsimonious relationship between the A–AS-level score and the GCSE score and present first the following basic model. The two scores are transformed to normality using normal scores for the whole sample:

$$\left.\begin{aligned} y_{ij} &= \beta_0 + \sum_{h=1}^{4} \beta_h x_{ij}^h + \beta_5 z_{ij} + u_j + e_{ij}, \\ u_j &\sim N(0, \sigma_u^2), \\ e_{ij} &\sim N(0, \sigma_e^2), \\ \rho &= \sigma_u^2(\sigma_u^2 + \sigma_e^2)^{-1} \end{aligned}\right\} \qquad (2)$$

where z_{ij} is a dummy variable coded 1 if the student is a girl and 0 if a boy and ρ is the intra-institution correlation and measures the extent of clustering of students within institutions. The fourth-degree polynomial is necessary to describe adequately the A–AS-level–GCSE relationship. Table 4 gives the parameter estimates for this model together with a model which makes no adjustment for GCSE score (C) and a model which adjusts for the 1992 mean institutional A–AS-score only (B). As Raudenbush and Willms (1995) point out, the unadjusted analyses cannot provide unbiased estimates of institutional effects; they are included here to illustrate the extent of the biases which result.

The unadjusted analysis has a clustering coefficient of 0.12 which is reduced considerably for the other two analyses. The effect of gender is that girls do significantly better than boys in terms of the unadjusted analyses, but worse when their GCSE results are taken into account, i.e. boys make more progress between GCSE and A-level than do girls. The student level variance is, of course, unchanged in the two analyses which differ only in terms of the 1992 institutional score, but the between-school variance is halved. When the GCSE score is introduced the between-student variance is halved and the between-institution variance reduced also.

We can estimate posterior means together with estimates of the standard errors of

TABLE 4
Basic analyses of A–AS-level scores

Parameter	*Estimates for the following models†:*		
	A	*B*	*C*
Fixed			
Intercept (β_0)	−0.09	−0.923	−0.076
GCSE (β_1)	0.80 (0.010)		
GCSE^2 (β_2)	0.11 (0.009)		
GCSE^3 (β_3)	−0.035 (0.004)		
GCSE^4 (β_4)	−0.005 (0.002)		
Girl (β_5)	−0.08 (0.01)	0.035 (0.014)	0.032 (0.015)
1992 score		0.059 (0.004)	
Random			
Level 2: σ_u^2	0.037 (0.004)	0.058 (0.006)	0.120 (0.011)
Level 1: σ_e^2	0.437 (0.004)	0.847 (0.008)	0.847 (0.008)
Intra-unit correlation: ρ	0.079	0.064	0.124

†Standard errors are given in parentheses.

these means for each institution (Goldstein, 1987, 1995), and we have done this for the analyses in Table 4 and then plotted these means against each other.

We see in Fig. 1 that the use of mean institutional scores as in the official league tables, but additionally adjusted for the gender difference, will result in quite different rankings for many institutions compared with the analysis which takes into account the student GCSE scores. Some institutions in which students appear to make good progress, i.e. after adjusting for GCSE performance, will be classified as having low rankings using the 'raw' mean and vice versa. This presents *prima facie* evidence therefore for the potential inequity and misleading nature of unadjusted league tables. The inclusion of the 1992 A–AS-score adjusts for the overall performance in the previous year but, as Fig. 2 shows, the rankings are also substantially different.

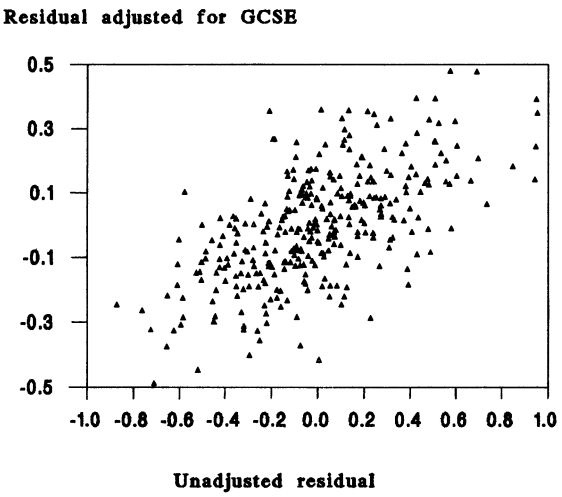

Fig. 1. Posterior means for analysis A *versus* analysis C in Table 4 (correlation, 0.62)

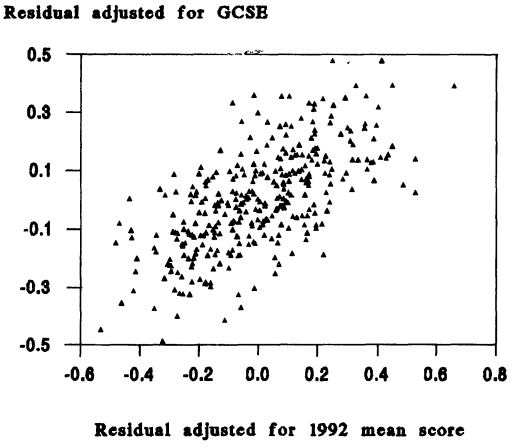

Fig. 2. Posterior means for analysis A *versus* analysis B in Table 4 (correlation, 0.65)

This indicates that changes in unadjusted raw scores over time are not a good predictor of institutional effectiveness as measured by its value-added estimate.

In the following set of analyses we explore further the relationship between GCSE scores and A–AS-level scores.

4. COMPLEX VARIATION BETWEEN INSTITUTIONS AND BETWEEN STUDENTS

It is of interest to establish whether the between-institution variation itself is a function of further factors since this would imply that the progress of students in a particular institution was also a function of those factors. In the present case we can study the possible effects of gender and GCSE score. We consider gender first and write

$$\left.\begin{aligned} y_{ij} &= \beta_{0j} + \sum_{h=1}^{4} \beta_h x_{ij}^h + \beta_{5j} z_{ij} + e_{ij}, \\ \beta_{0j} &\sim N(\beta_0, \sigma_{u0}^2), \\ \beta_{5j} &\sim N(\beta_5, \sigma_{u5}^2), \\ \operatorname{cov}(\beta_{0j}, \beta_{5j}) &= \sigma_{u05}. \end{aligned}\right\} \tag{3}$$

For notational purposes we have incorporated the level 2 institutional effect into the intercept. Table 5 shows the results of fitting this model together with interactions in the 'fixed' part of the model between gender and GCSE score. It also introduces complex variation at level 1, i.e. between students, by modelling this variance as a function of gender, i.e. allowing separate variances for males and females so that we write

TABLE 5
Analyses of gender differences

Parameter	*Estimates for the following models*†*:*	
	A	*B*
Fixed		
Intercept (β_0)	−0.10	−0.10
GCSE (β_1)	0.84 (0.011)	0.84 (0.011)
GCSE^2 (β_2)	0.12 (0.010)	0.12 (0.010)
GCSE^3 (β_3)	−0.035 (0.004)	−0.035 (0.004)
GCSE^4 (β_4)	−0.005 (0.002)	−0.005 (0.002)
Girls (β_5)	−0.071 (0.01)	−0.072 (0.02)
GCSE×girls ($\beta_{1,5}$)	−0.074 (0.01)	−0.074 (0.01)
GCSE^2×girls ($\beta_{2,5}$)	−0.013 (0.008)	−0.013 (0.008)
Random, level 2		
σ^2_{u0}	0.041 (0.005)	0.039 (0.005)
σ_{u05}	−0.008 (0.003)	−0.006 (0.003)
σ^2_{u5}	0.012 (0.004)	0.012 (0.004)
Random, level 1		
σ^2_{e0}	0.433 (0.004)	0.475 (0.007)
σ_{e05}		−0.04 (0.004)

†Standard errors are given in parentheses.

$$\left.\begin{aligned} e_{ij} &= e_{0ij} + e_{5ij}x_{5ij}, \\ e_{ij} &\sim N(0, \sigma^2_{e0} + 2\sigma_{e05}x_{5ij}), \\ \mathrm{var}(e_{0ij}) &= \sigma^2_{e0}, \\ \mathrm{cov}(e_{0ij}, e_{5ij}) &= \sigma_{e05} \end{aligned}\right\} \tag{4}$$

where x_{5ij} is coded 1 for females and 0 for males, so that $2\sigma_{e05}$ is the difference between the variances for females and males.

There is an interaction between gender and GCSE score up to the second-order term whereby the slope of the relationship between GCSE score and A-level is less for girls. In addition the between-institution variance for females is 0.038 compared with 0.041 for males and the, level 1, between-female variance is 0.40 compared with 0.48 for males. Strictly, therefore, when comparing institutions we should be concerned with the differences for both males and females. In fact the correlation between the estimated residuals for males and females for those schools with both genders is 0.94.

The likelihood ratio χ^2 for testing model A against the variance components model with a single variance at both levels is 18.5 with 2 degrees of freedom ($P<0.0001$) and the likelihood ratio χ^2 for testing model B against model A is 89.3 ($P<0.0001$). A large sample test, based on the estimated covariance matrix of the fixed coefficients, for the two interaction terms of gender with GCSE gives a value of 57.5 compared with a likelihood ratio test statistic of 57.4, both on 2 degrees of freedom. It is generally adequate in the fixed part of the model to use the former test statistic, or for a single parameter the standard error estimate.

We now extend the model by studying the dependence of the variation at both levels on the GCSE score. One way of incorporating such dependence is by allowing the coefficients of the GCSE score to vary randomly at level 2 and to make the level 1 variance a linear, quadratic etc. function of GCSE score (Goldstein, 1995). Another way is to group the GCSE score so that there is variance homogeneity within but not between groups. This allows results to be presented directly in terms of institutional differences for students within each GCSE score band which has interpretational advantages. We have therefore explored the data by carrying out analyses using different groupings of the GCSE score. These analyses suggest that three groups provide an adequate description, with boundaries at the lower quartile, median and upper quartile. This extended model can now be written as

$$y_{ij} = \beta_0 + \sum_{h=1}^{4} \beta_h x_{ij}^h + \beta_{5j} z_{ij} + \beta_6 x_{ij} z_{ij} + \beta_7 x_{ij}^2 z_{ij} + u_{1j} w_{1ij} + u_{2j} w_{2ij} + u_{3j} w_{3ij} + u_{5j} z_{ij} + e_{ij} \tag{5}$$

where the w_{hij} for $h = 1, 2, 3$ are dummy (0, 1) variables for the three GCSE groups. The level 1 variation is given by

$$\text{var}(e_{ij}) = \sigma_{e0}^2 + 2\sigma_{e05} z_{ij} + 2\sigma_{e02} w_{2ij} + 2\sigma_{e03} w_{3ij}.$$

This variance is an additive function of parameters for GCSE group and gender, with one category in each factor omitted. In general we could consider an interactive function where a separate parameter for each of the six combinations of gender by GCSE group was specified. In the present case this does improve the fit ($\chi_2^2 = 31.5$, $P < 0.0001$) and the full level 1 variance function is given in Table 6. We see an interesting reversal with the variation being greater for males in the lowest GCSE group (1) but considerably higher for females in the highest GCSE group (3). The remaining parameters of the model are little changed.

The incorporation of the three GCSE groups at level 2 is highly significant with a likelihood ratio χ^2 of 129.6 with 6 degrees of freedom. The addition of heterogeneous variance based on these groups at level 1 yields a likelihood ratio χ^2 of 11.1 with 2 degrees of freedom ($P = 0.004$) with a slightly higher between-student variance for the middle GCSE group.

Table 7 and Table 8 show the result of fitting model (5). The fixed part of the model changes little but we now have complex variation between institutions with only moderate correlations between the effects for the GCSE groups. To illustrate this Fig. 3 plots the estimated residuals for the lowest GCSE group against those for the highest group, for boys. These results show that for some institutions there are substantial differences in the average progress of the most able and least able

TABLE 6
Level 1 variances

GCSE group	*Variance for males*	*Variance for females*
1	0.502	0.345
2	0.476	0.559
3	0.425	0.613

TABLE 7
Complex level 1 variation for GCSE groups and gender

Parameter	*Estimate (standard error)*
Fixed	
Intercept (β_0)	−0.09
GCSE (β_1)	0.83 (0.013)
GCSE^2 (β_2)	0.11 (0.010)
GCSE^3 (β_3)	−0.033 (0.004)
GCSE^4 (β_4)	−0.004 (0.002)
Girls (β_5)	−0.064 (0.01)
GCSE×girls ($\beta_{1,5}$)	−0.077 (0.01)
GCSE^2×girls ($\beta_{2,5}$)	−0.016 (0.008)
Random, level 1	
σ^2_{e0}	0.451 (0.010)
σ_{e02}	0.015 (0.005)
σ_{e03}	0.003 (0.006)
σ_{e05}	−0.041 (0.004)

TABLE 8
Level 2 correlations of between-institution effects: variances on the diagonal

	GCSE group 1	*GCSE group 2*	*GCSE group 3*	*Gender difference*
GCSE group 1	0.068			
GCSE group 2	0.71	0.042		
GCSE group 3	0.43	0.85	0.043	
Gender difference	−0.52	−0.29	0.01	0.014

and that a single 'effectiveness' measure may mask important within-institution differences.

The correlation between these residual estimates is 0.56 which compares with 0.43 from the parameter estimates in Table 8 and reflects the fact that these residuals are shrunken regression estimates. For comparison with Fig. 1, Fig. 4 shows the residuals for the middle GCSE group plotted against the raw score, with a correlation of 0.59.

To check the model assumptions we have studied plots of the residuals and in Fig. 5 we show one of these, for the lowest GCSE group for boys, which shows the greatest departure from normality, although this is not substantial. Plots of residuals by predicted values did not show evidence for variance heterogeneity.

5. NUMBER OF EXAMINATIONS TAKEN, AGE, STATUS AND TYPE OF INSTITUTION

The restriction of the GCSE score to the eight best grades will tend to underadjust for the very high achieving students and three-quarters of the students actually have nine or more GCSE grades. We have therefore included the number of GCSE examinations taken as a further explanatory variable. We also carried out an analysis using the average GCSE grade, but this made little difference since less than 3% took

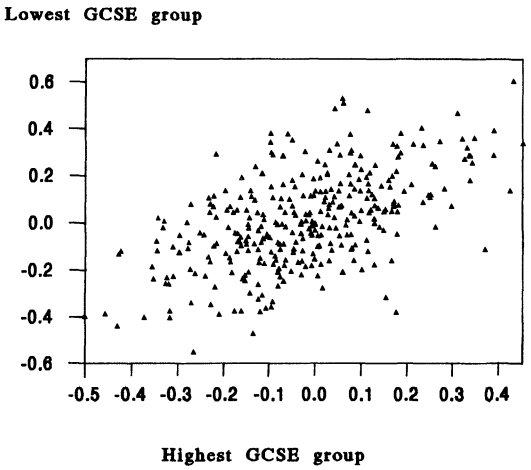

Fig. 3. Estimates of residuals for GCSE groups for each institution for boys

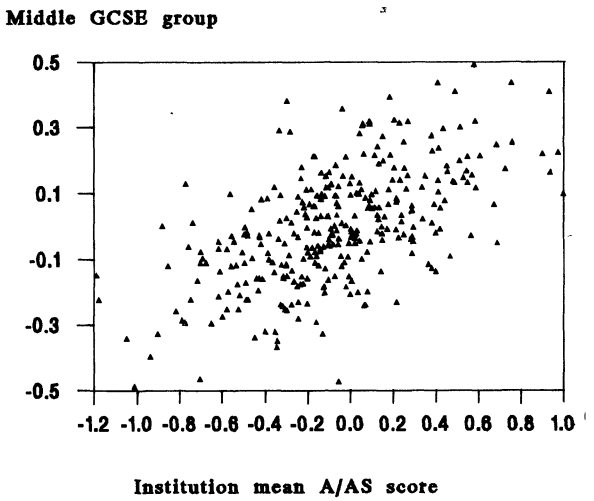

Fig. 4. Estimates of residuals compared with raw A–AS-score for boys

fewer than eight GCSE examinations. We have also included the age of the student and the type and status of the institution. Table 9 presents the results for the analysis which includes these additional variables. We also studied the effect of institutional status (local education authority maintained, voluntary, grant maintained, private, other) but there were no significant differences ($\chi^2 = 5.0$, 4 degrees of freedom). We do not present the level 1 and level 2 random parameters since these are little altered from Table 7 and Table 8.

Some care is needed in interpreting these results since the sample is not representative of all institutions and there may be biases reflected in the results. Nevertheless, the introduction of the number of GCSEs taken shows that, between eight and 10, the fewer GCSEs taken, over the eight best GCSE results, the higher the predicted A-level score. This may be a reflection of entry policies where some of the

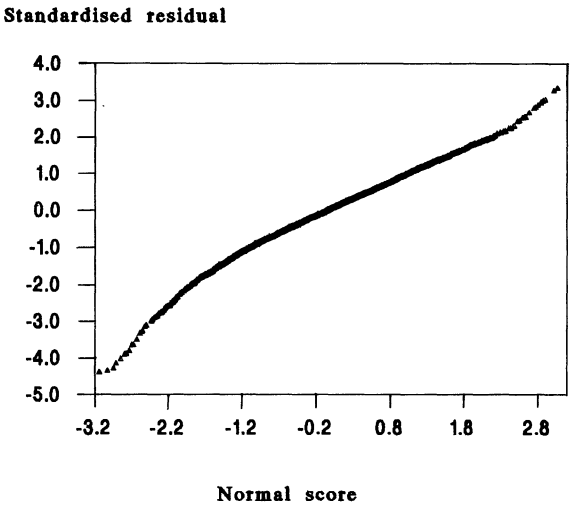

Fig. 5. Lowest GCSE group residuals by normal score

highest achieving students are not entered for a large number of GCSE subjects. There is a small decrease in predicted A-level score with age and the sixth-form college results are somewhat worse than those of other institutions, with the comprehensive schools' results also somewhat worse than those from the selective grammar schools. The girls on average make less progress than the boys and this is most noticeable for those girls with higher GCSE scores. This contrasts with the overall slightly better performance of girls shown in Table 4, model C.

The level 2 residuals, the institutional effects, from this analysis correlate highly with those from the previous analysis, namely 0.91, 0.92 and 0.94 for the lowest, middle and highest GCSE groups.

TABLE 9

A–AS-level score related to GCSE, gender, school type, number of GCSEs and age

Parameter	*Estimate (standard error)*
Intercept	3.02
GCSE	0.83 (0.12)
$GCSE^2$	0.13 (0.01)
$GCSE^3$	−0.029 (0.004)
$GCSE^4$	−0.010 (0.002)
Girls	−0.066 (0.015)
Girls×GCSE group 2	−0.075 (0.010)
Girls×GCSE group 3	−0.020 (0.008)
No. of GCSEs	−0.425 (0.058)
$(\text{No. of GCSEs})^2$	−0.019 (0.003)
Age (years)	−0.053 (0.013)
Comprehensive−6th-form college	0.091 (0.084)
Grammar−6th-form college	0.188 (0.086)
Other (excluding further education and tertiary)−6th-form college	0.180 (0.106)

6. UNCERTAINTY INTERVALS

The residual estimates u_{hj}, $h = 1, 2, 3$, are sample estimates and have errors attached to them. We can obtain estimates of these standard errors which will depend on the number of students in the institution and the between and within-institution variation. If we wish to make comparisons between sets of institutions then it will be necessary to take account of this uncertainty. The usual formulae for these standard errors will tend to produce underestimates because they use estimated parameter values (Goldstein, 1995). In the present case, however, the large number of institutions in the sample means that this bias is negligible. Likewise, the fitting procedure generally will result in non-zero covariances between residual estimates for different institutions; these are small and have been ignored.

In the simplest case, suppose that we wish to compare two institutions. If we assume that the estimates of residuals are independent a large sample test can be carried out using the estimated standard errors and assuming normality. An equivalent procedure is to construct a normal confidence interval about each estimate and to judge separation as significant if the intervals do not overlap. Thus, for a 95% test we would construct confidence intervals ± 1.40 times the standard errors. If now we envisage that all the residuals are available and that on average each pair of institutions will be compared the same number of times, we can construct an interval for each institution so that, on average, the type 1 error is at the required level. Goldstein and Healy (1995) provide details of such a procedure. The procedure can be extended to consider comparisons of triplets etc. In practice, if comparisons are being made by students and parents, then the procedure can be extended by weighting each institution according to the number of times that it enters a comparison. The assumption of institutions being compared the same number of times in pairs will result in the narrowest set of interval widths.

Fig. 6 shows a set of 95% intervals for a random sample of 75 institutions. It is clear that for the majority most of the pairwise comparisons will not allow the institutions to be separated. Remembering that this gives the most 'conservative' picture which will tend to overestimate differences, it suggests that the use of examination results for comparative purposes will be rather uninformative. This will be particularly so where we wish to make comparisons for individual subjects where at A-level the numbers of students often will be very small indeed.

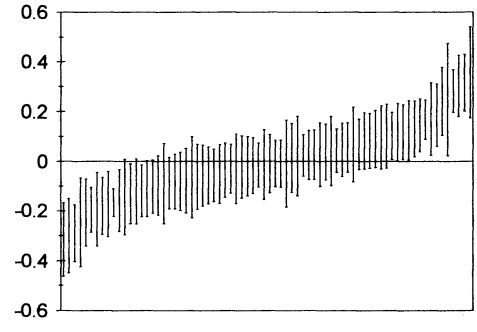

Fig. 6. Pairwise 95% uncertainty intervals for the middle GCSE group

7. DISCUSSION

The principal purpose of this paper has been to explore the use of examination results as indicators for comparing institutions. We have argued that the use of raw, unadjusted, results is invalid and have demonstrated that even when adjustments are possible the comparison of individual institutions is imprecise and that in most cases statistical separation is impossible. Our conclusions will also be applicable generally to other stages of the education system, e.g. to GCSE examinations at 16 years (Goldstein *et al.*, 1993; Thomas *et al.*, 1994, 1996) and to national assessment test scores (Thomas and Nuttall, 1992).

Although it was not the primary purpose of the analysis, we have also carried out comparisons between male and female students, for students of different ages and between types of institution. This suggests that there may be some interesting differences, especially with respect to gender differences, but the present results must remain tentative given the low response rate achieved. Our conclusions concerning the validity of institutional comparisons, however, are unlikely to be influenced markedly by sample bias. As demonstrated in Table 3, the institutions which supplied information had A-level results that were slightly lower than average. It is conceivable that these institutions were also those with, on average, greater discrepancies between adjusted and unadjusted A-level scores, but it is difficult to think of a mechanism whereby this could have occurred systematically. Furthermore, our findings are consistent with those from other studies. A further issue concerns examination policies. Students may be entered for examinations other than A- or AS-level and some institutions may have policies of encouraging weaker students to do so. Without further information on entry policies, however, we can provide no insights into the effects of such policies.

Where it is possible to study institutions over time, the time trends will provide further information which can be of interest, although relatively long time periods will be required. In addition, for A–AS-level results, the information for judgment will become available for use typically 3 years after the cohort being studied entered their institutions. In the meantime those institutions may have changed and we may need to rely on more qualitative local judgments about such changes. This makes the use of this kind of data for choice purposes, whether adjusted or not, quite problematic. It is indeed a feature of most indicator systems when used in such a way and implies that the aim in the *Parents' Charter* of using examination results for school comparisons cannot be achieved: there is no simple method of comparison which can achieve fair and accurate comparisons between institutions. It follows that the publication of league tables without a clear statement of their limitations is both misleading and scientifically unjustifiable. This point is also made in a report by the School Curriculum and Assessment Authority on value-added measures (School Curriculum and Assessment Authority (1994), chapter 3, section 3(b)).

However, as a device for monitoring education and for attempting to explain institutional differences, the methods of this paper, suitably extended to a large random sample of institutions, and with the collection of further student level and institution level data, could be used effectively.

It is also possible to argue, for those institutions at the extremes, that our procedures can be used as a screening device to identify possible problems. Used sensitively, by those charged with supporting rather than merely judging schools,

such a screening procedure could have value among other sources of information.

ACKNOWLEDGEMENTS

This paper owes much to the inspiration and dedication of the late Professor Desmond Nuttall who first proposed the survey of A-level results. We wish also to thank Stephen Bates and James Meikle of *The Guardian* for their support and co-operation which has successfully introduced to a wide public some important issues in the interpretation of statistical data. We also wish to thank the following for their comments on a draft: Min Yang, Jon Rasbash, Pam Sammons, Ian Plewis, Steve Raudenbush and Doug Willms. This work was partly supported by the Economic and Social Research Council under Analysis of Large and Complex Datasets Programme.

REFERENCES

Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies (with discussion). *J. R. Statist. Soc.* A, **149**, 1–42.

Department of Education and Science (1991) *The Parents' Charter*. London: Her Majesty's Stationery Office.

———(1993) *The Guardian, Education Supplement*, Nov. 30th.

Goldstein, H. (1987) *Multilevel Models in Social and Educational Research*. London: Arnold.

———(1995) *Multilevel Statistical Models*, 2nd edn. London: Arnold.

Goldstein, H. and Healy, M. J. R. (1995) The graphical presentation of a collection of means. *J. R. Statist. Soc.* A, **158**, 175–177.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. L. and Thomas, S. (1993) A multilevel analysis of school examination results. *Oxf. Rev. Educ.*, **19**, 425–433.

Guardian (1992) *The Guardian, Education Supplement*, Oct. 20th.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *J. R. Statist. Soc.* B, **34**, 1–41.

Raudenbush, S. W. and Willms, J. D. (1995) The estimation of school effects. *J. Educ. Behav. Statist.*, to be published.

Scheerens, J. (1992) *Effective Schooling*. London: Cassell.

School Curriculum and Assessment Authority (1994) *Value-added Performance Indicators for Schools*. London: School Curriculum and Assessment Authority.

Thomas, S. and Mortimore, P. (1996) Value added analysis of 1993 GCSE results. *Res. Pap. Educ.*, to be published.

Thomas, S. and Nuttall, D. L. (1992) An analysis of 1991 Key Stage 1 results in Dorset: a multilevel analysis of English, Mathematics and Science subject level scores. *Br. J. Curr. Assessmnt*, **3**, 18–20.

Thomas, S., Pan, H. and Goldstein, H. (1994) *The Analysis of 1992 GCSE Results*. London: Association of Metropolitan Authorities.