Methods of aggregating assessment results to predict future examination performance

John F Bell
Research and Evaluation Division
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge
CB1 2EU
☎01223 553849
Fax: 01223 552700
💻 bell.j@ucles.org.uk

Key words: Multilevel model, GCSE, A-level

Abstract

One of the problems in measuring progress in educational research is the choice of variables to include in the study. In this paper, the issue of creating summaries of GCSE results with the objective of predicting A-level results is investigated. The advantages and disadvantages of using measures such as total GCSE, mean GCSE, or other combinations of GCSE is considered. The effect of these measures is described by applying them to candidates from a sample of examinations.

It was found that the mean GCSE was not the best predictor of success for individual A-level subjects. For predicting high attainment at A-level, the best predictor proved to be the sum of the square roots of the best five GCSE grades. This measure rewards a steady performance on a limited range of subjects. For lower levels of A-level attainment, a different measure was found to be optimal. This was the mean of the squares of the GCSE points for all subjects sat by each candidate. This measure rewards erratic performances.

Disclaimer

The opinions expressed in this paper are those of the author and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate.

Introduction

In quantitative school effectiveness studies, information about prior attainment is often used to assess the value-added by the individual schools (e.g. DfE, 1995, O'Donoghue et al., 1997, Yang and Woodhouse, 2001). In most cases, the relative progress is measured because a general measure of prior attainment is used. In many studies carried out in the UK, this has involved aggregating public examination results. Aggregation is used because pupils are given a choice in the subjects they study which makes it difficult to use a profile of separate scores. In this paper, methods of combining the GCSE (General Certificate of Secondary Education) grades awarded to candidates will be investigated. These aggregated measures will be used to investigate the effect of predicting performance in A-level examinations. Although this is a specific example of aggregation, some of the issues are of interests in other circumstances.

In England, Wales, and Northern Ireland, the GCSE is an examination that is mostly taken by sixteen-year-olds and A-level examinations are usually taken by eighteen-year-olds. There are separate examinations for each subject and these examinations are administered by five awarding bodies. The aggregation of GCSE scores is not a simple issue because of the diversity of subjects taken by year 11 pupils (16 year olds). Despite the introduction of the National Curriculum in England there is still a considerable degree of flexibility in the choice of subjects (Bell, 1998, in press). A particularly important example of this diversity relates to the core subject of science. Although pupils are required to study science, there is flexibility in the amount of science that they can study (i.e. single award, double award and the three separate sciences). This choice has obvious implications for the uptake of other subjects.

The data used in this paper were extracted from two databases: the 1999 16+ database and the 1999 16+/18+ database. These databases are used for the compilation of performance (league) tables. The first database includes all the GCSE results for candidates in English schools born between the 1st September 1982 and the 31st August 1983. The second database contains all the GCSE, AS and A-level results for candidates in English schools born between 1st September 1980 and 31st August 1981. The analyses described in this paper were carried out on several samples extracted from these databases. This was done because the size of the data sets means that it is impractical to carry out the exploratory work described in the paper on the whole data set (e.g., in the 1999 16+ database there are approximately 800,000 candidates entered for more than 5,000,000 examinations). In the first sample, drawn from the 16+ database, the selection was based on the centres in which candidates were entered for GCSEs and it was chosen to investigate the effect of the different measures on all GCSE centres and candidates. The other samples were drawn from the 16+/18+ database. The first A-level sample was used to investigate how well the GCSE summary measures predicted A-level. In addition to these samples, samples of centres were selected for English and mathematics (other subjects are considered in Bell, 2000).

This report is divided into the following sections:

- Scoring systems for aggregate measures of GCSE performance
- Aggregate measures of GCSE performance and their properties
- Comparisons of the different aggregate measures
- Using aggregate measures to predict the A-level points score
- Using aggregate measures to predict results for individual A-level subjects.

Finally the report concludes with a discussion of some of the issues raised by this research.

Scoring systems for aggregate measures of GCSE performance

Candidates taking GCSE examinations are awarded grades ranging from A* to G for each subject. In most research involving aggregated GCSE results, the GCSE grades are converted into scores as in the second column of Table 1 (or sometimes the reverse order is used). Then either of the two simplest aggregations,

mean or total, are used. Gray et al. (1999) note that no single summary examination statistic can capture the dimensions of school performance. Gray, Goldstein, and Jesson, (1996) used the total GCSE score in a study of school effectiveness. They made the assumption that the highly correlated nature of the different examination measures meant that they would have obtained similar findings with other measures. They did not, however, rule out the possibility that they had favoured some schools more than others.

Table 1: Conversion of grades into points

Grade	Usual J=1	Steady J=1/2	Erratic J=2
A*	8	2.83	64
A	7	2.65	49
B	6	2.45	36
C	5	2.24	25
D	4	2.00	16
E	3	1.73	9
F	2	1.41	4
G	1	1.00	1
U	0	0.00	0

The conventional points system in Table 1 assumes that there is a linear relation between grades and the underlying attainment. This might not be case. It is possible to investigate other relationships which tend to reward steady or erratic performances (Wood, 1991). Depending on how the scores are aggregated it is possible to reward steady or erratic performances. To investigate this, it is necessary to consider the following power series:

$$C = \sum \alpha_i^j$$

where a_i is the grade score for examination *i* and *j* is a constant. The choice of *j* determines whether erratic or steady performances are rewarded. This can be illustrated by considering the following example. Consider the results of two candidates who have taken five examinations and obtained the following grade scores:

Candidate A: D, D, D, D, D (A steady candidate)
Candidate B: U, A*, A*, D, U (An erratic candidate)

By setting *j*=1, neither erratic nor steady candidates are favoured, i.e.

Candidate A scores (4+4+4+4+4)=20
Candidate B scores (0+8+8+4+0)=20

Taking a value $j < 1$ will favour steady candidates over erratic candidates. For example, take the value j=1/2 (i.e. square roots) gives the following:

Candidate A: (2+2+2+2+2)=10
Candidate B: $(0+\sqrt{8}+\sqrt{8}+2+0)=7.66$.

Taking values greater than 1 rewards erratic performances. For example, taking j=2 (i.e. square the scores) gives the following:

Candidate A: (16+16+16+16+16)=80
Candidate B: (0+64+64+16+0)=144

In the usual transformation, a grade A* is worth twice as much as a grade D. When the square root transformation is applied, a grade A* is only worth 1.41 times as much as a grade D. However, the effect of squaring the points is even more dramatic. A grade A* is worth 4 times as much as a grade D. This means that the amount of compensation allowed for a good performance on a subject varies with the scoring system. In the remainder of this paper, all three points systems given in Table 1 will be considered.

Aggregate measures of GCSE performance and their properties

There are many different aggregate measures that could be used to summarise GCSE performance. In this section, a number of different measures have been defined and their potential advantages and disadvantages are considered. The aggregate measures of GCSE performance used in this paper are given in Table 2. Each of the measures has been given an identification code that is used in other figures and tables in this paper. Two of the measures use k to denote a value that could be varied. For example, the measure based on the best 5 GCSEs would be denoted as BE5. The measure 'At least 5 Cs at GCSE' has been included because it is used in the published school performance tables. Although this measure has a long history dating back to the change from the School Certificate to O-level, it has only been included for completeness and is not seriously considered as a suitable predictor of the grades at A-level (the vast majority of A-level candidates satisfy this requirement).

Table 2: Aggregate measures of GCSE performance

ID	Name	Description
MN	Mean GCSE score	Take the mean GCSE scores for each candidate
TOT	Total GCSE score	Add up all the GCSE scores for each candidate
MN*k*	Mean GCSE with minimum N fixed	Divide the total scores by n if N>*k* otherwise by *k*. E.g. if *k* was set at 5, AAAAB would lead to a score of 6.8 and AAAA would score 5.6
BE*k*	Best k GCSE scores	Take the best k GCSE scores
P5C	At least 5 grade Cs	Used in performance tables

For each of the measures in Table 2, it is possible to calculate the measure for each of the scoring systems in Table 1. When the square roots are used, an upper case R is added to the end of the ID and when the squares are used then an upper case S is added. If the square roots scores were used, the measure 'best 5 GCSE scores' would be identified as BE5R.

The reason for developing these measures is that they each have advantages and disadvantages (although whether a characteristic is an advantage or a disadvantage can be subjective). They also favour particular profiles in different ways. This information relating to the characteristics of different measures is given in Table 3.

Table 3: Some characteristics of different measures of GCSE performance

ID	Characteristics
MN	Highest scoring candidates tend to be the most consistent Resits are penalised AAAAB (34 points giving a mean of 6.8 points) is worse than AAAA (28 points giving a mean of 7.0 points) Still influenced by school policy on choice of subjects.
TOT	Candidates who sit very few GCSEs tend to obtain lower scores than ones who do more. Highest scoring candidates tend to be those who take the most GCSEs. Number of GCSEs taken may be a policy decision of the centre Candidates gain by resits
MN*k*	A compromise between MN and TOT Choice of *k* is arbitrary Favours candidates who take exactly *k* examinations
BE*k*	School/national policies which require pupils to take certain subjects have less influence Choice of *k* is arbitrary
P5C	Does not discriminate well or at all for high attaining candidates

It is worth noting in passing that the choice of measure would lead to different strategies for maximising school performance. If TOT is used then the best strategy is to enter candidates for more GCSEs (Gray et al. (1999) pointed out that one result of the introduction of performance tables in 1992 was to increase the average number of GCSE examinations sat in the schools in their study). If MN is used then a better strategy would be to reduce the number of entries per candidate.

For the purposes of this paper, two values of k have been considered. Firstly, the value of five has been chosen because of the use of five passes at C in performance tables. Secondly, the value of nine has been chosen because it is the modal number of GCSEs taken by year 11 pupils.

Comparisons of the different aggregate measures

For all the samples described in the introduction, the aggregate measures described in the previous section were calculated. For the purposes of these analyses, only those GCSEs taken by the candidates up to and including year 11 were used. The aggregate measures were standardised so that they all had a mean of zero and a standard deviation of 1. This was to investigate the effect of the different measures on the mean aggregate score for the sample of centres considered in this study.

The relationships between the measures are investigated by considering the first sample, which included the full range of GCSE attainment. This sample consisted of all the candidates (9,060) from 50 GCSE centres selected at random from the database. These relationships are illustrated in Figures 1 and 2, which are scatterplot matrices. The point of the plots is simple. When there are many variables to plot against each other in scatterplots, it is logical to arrange the plots in rows and columns using a common vertical scale for all plots within a row (and a common horizontal scale within columns). In the diagonal cells, a frequency polygon indicating the distribution of each measure is given. The first scatterplot matrix (Figure 1) considers the relationships between the three scoring methods by plotting MN, MNR and MNS. The effect of using the square root transformation is that candidates with high scores for MN tend to have lower scores for MNR. Taking the square of the score has the reverse effect. This same general pattern occurs for the other measures.

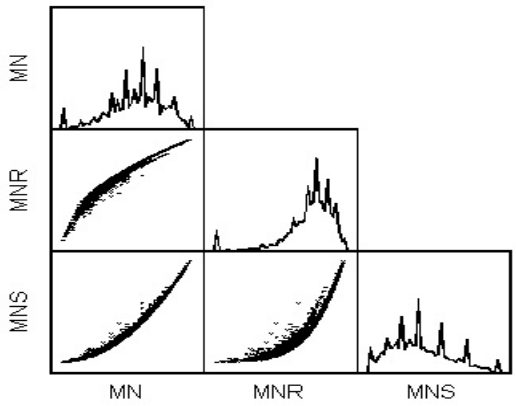

Figure 1: Scatterplot matrix for measures MN, MNR and MNS

Figure 2 is a scatterplot matrix for the five measures MN, TOT, MN9, BE5 and BE9. It is clear that the values of MN can be very different for some candidates (ones who did well but only attempted a small number of GCSEs). The two measures that exhibit the greatest similarity are MN9 and BE9. As expected, BE9 and TOT tend to differ more for higher levels of performance because more able candidates tend to take more GCSEs.

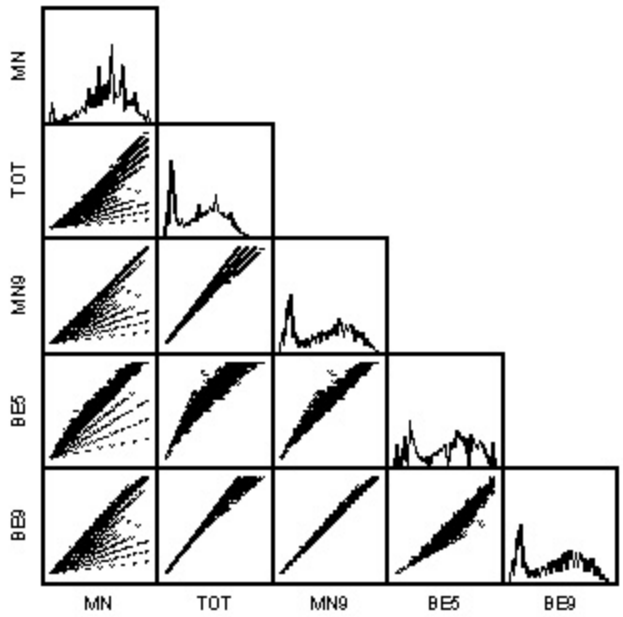

Figure 2: Scatterplot matrix for the measures MN, TOT, MN9, BE5 and BE9

Using the standardised aggregate measures as dependent variables, separate multilevel models were fitted to partition the variation into pupil-level and centre-level variation for the first sample. In addition to the aggregate measures that were later used to predict A-level performance, the measure, at least 5 grade Cs – P5C, was included in the analysis. This measure is not used in the remainder of the report because it is not an effective predictor of A-level performance. The results of these analyses have been presented in Table 4. As expected, the intercepts for all the standardised measures are small and near zero. The amount of centre variation is lowest for P5C. This is not surprising because this measure only considers differences toward the middle of the attainment range. For example, selective and independent schools usually have similar very high values of P5C but could differ substantially on the other measures.

Table 4: Results of multilevel models for sample 1

	Intercept		Centre		Candidates		
Measure	Est.	s.e.	Est.	s.e.	Est.	s.e.	% of variation
MN	-0.02	0.09	0.39	0.08	0.74	0.01	35
MNR	-0.03	0.09	0.32	0.07	0.79	0.01	29
MNS	-0.01	0.09	0.40	0.08	0.71	0.01	36
TOT	0.03	0.09	0.41	0.08	0.66	0.01	38
TOTR	0.03	0.09	0.41	0.09	0.64	0.01	39
TOTS	0.04	0.09	0.40	0.08	0.69	0.01	37
MN9	0.04	0.09	0.41	0.09	0.66	0.01	38
MN9R	0.03	0.09	0.42	0.09	0.64	0.01	40
MN9S	0.05	0.09	0.40	0.08	0.69	0.01	37
BE5	0.04	0.09	0.40	0.08	0.67	0.01	37
BE5R	0.03	0.09	0.40	0.08	0.64	0.01	38
BE5S	0.04	0.09	0.39	0.08	0.69	0.01	36
BE9	0.04	0.09	0.41	0.08	0.66	0.01	38
BE9R	0.03	0.09	0.40	0.08	0.64	0.01	38
BE9S	0.04	0.09	0.40	0.08	0.69	0.01	37
P5C	0.04	0.08	0.27	0.05	0.78	0.01	26

The mean scores for each centre on each of the aggregated measures are displayed on Figure 3. It is a parallel coordinates plot. Each vertical line is an axis of a particular aggregate measure. The mean aggregate measures of the centres have been plotted on each of the axes. The points for each centre have been joined by line segments. The point of this figure is that if the lines do not cross then the rank order of the centres would not change. This plot indicates that the measures based on mean GCSE are substantially different from the other measures. The fact that the lines in this plot cross means that the choice of measure would affect the position of the centres if a performance table was produced. There is a small group of centres at the bottom of the plot that only enter candidates for a small number of GCSEs per candidate. It should be recognised that the sample was drawn from all centres that offered GCSEs to Y11 pupils in 1999 and includes some centres that would not normally be included in performance tables. Since there is variation in the centre mean scores for the various measures then it is likely that the different measures might explain different amounts of centre level variation in the multilevel models described in the later sections of this report.

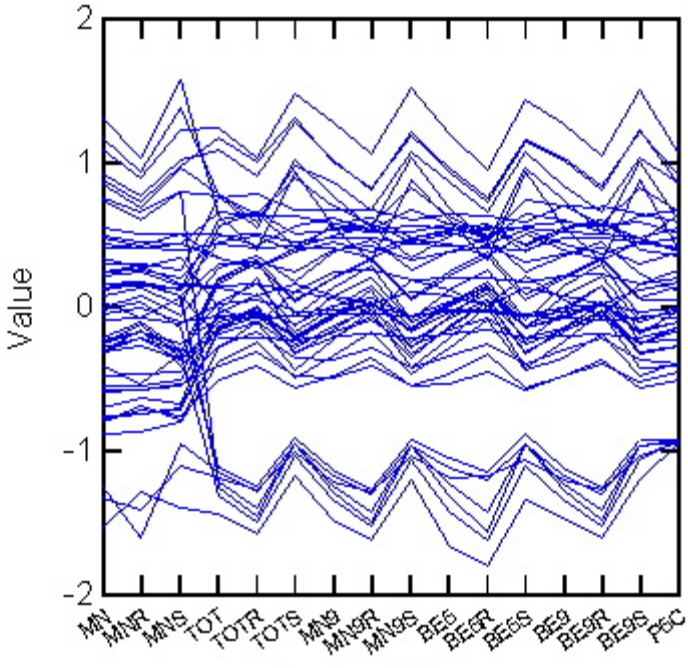

Figure 3: Parallel co-ordinate plot for the means of the standardised measures

The results of these analyses suggest that all the measures are sufficiently different from each other for them to be included in further analysis. All the measures other than P5C will be included in the subsequent multilevel models described in the remaining sections of this paper.

Using aggregate measures to predict the A-level points score

Like GCSE, A-level is awarded in grades, but for A level pass grades range from A to E. At the time of this paper, candidates usually entered three subjects. In this section, the results will be presented of a series of analyses that involved using the aggregate measures to predict total A-level points scores. A-level points were obtained by using the following system: A – 10, B – 8, C – 6, D – 4, E – 2, U/N – 0.

Other research into the relationship between GCSE and A-level has indicated that the relationship between the two measures is non-linear. The relationship was investigated using the second sample which consisted of all the pupils taking both A-levels and GCSEs (2,057 pupils) from a randomly selected sample of 50 A-level centres. The following multilevel model was fitted:

$$y_{ij} = \beta_0 + \beta_1 meas_{ij} + \beta_2 meas_{ij}^2 + \gamma_{0i} + \varepsilon_{0ij}$$

where y_{ij} is the A-level points score for candidate *j* from centre *i*,

β_0 is a constant,

$meas_{ij}$ is the aggregate measure of GCSE performance,

β_1, β_2 are slope parameters,

γ_{0i} is a random slope parameter,

and ε_{0ij} is a random error term.

All the aggregate measures of GCSE performance were standardised with mean zero and variance one for only those candidates with both GCSE and A-level results. This means that the value of the intercepts in the multilevel models can be interpreted as the A-level performance of candidates at the mean of the aggregate measure. For most of the measures, this is a total A-level score of approximately 16 points (e.g, 2 Cs and 1 D). The results for the multilevel models are presented in Table 5. The differences in the amount of reduction in candidate and centre level error are not large for most of the measures. However, the best measure seems to be MNS followed by BE5.

For the most powerful models used here, the standard deviation of the centre parameters is 2.3 A-level points (or just over 1 grade in one of three subjects). It is probable that a more sophisticated model with more explanatory variables would explain more of this centre level variation. Although any difference between centres could mean that the choice of centres has an effect on an individual candidate's life chances, after accounting for prior achievement the amount of centre level variation is not large.

Table 5: Results for the multilevel models with total A-level points score

Name	Int.	s.e.	Meas.	s.e.	Meas^2	s.e.	Centre	s.e.	Cand.	s.e.	Total
None	16.5	0.8	-	-	-	-	26.1	6.6	95.5	3.0	123.6
MN	16.4	0.4	8.5	0.2	1.2	0.1	5.5	1.5	41.1	1.3	46.6
MNR	16.0	0.4	8.9	0.2	1.6	0.1	6.0	1.6	43.0	1.4	49.0
MNS	**17.2**	**0.4**	**8.4**	**0.2**	**0.5**	**0.1**	**5.3**	**1.5**	**40.1**	**1.3**	**45.4**
TOT	15.7	0.5	9.8	0.2	2.0	0.1	9.2	2.4	48.9	1.5	58.1
TOTR	15.4	0.7	10.9	0.3	2.1	0.1	16.9	4.4	62.4	2.0	79.3
TOTS	16.7	0.4	8.3	0.2	1.1	0.1	5.9	1.7	44.0	1.4	50.4
MN9	15.4	0.7	10.9	0.2	2.3	0.1	5.9	1.7	42.9	1.4	48.8
MN9R	14.9	0.5	15.0	0.3	2.8	0.1	8.0	2.4	48.6	1.5	56.6
MN9S	16.4	0.4	8.5	0.2	1.4	0.1	6.3	1.7	42.3	1.3	48.6
BE5	15.4	0.4	11.6	0.2	2.3	0.1	5.1	1.4	41.4	1.3	46.5
BE5R	14.5	0.4	18.0	0.4	3.1	0.1	5.3	1.5	42.7	1.3	48.0
BE5S	16.0	0.4	8.7	0.2	1.7	0.1	5.5	1.5	42.0	1.3	47.5
BE9	16.4	0.4	8.5	0.2	1.4	0.1	5.6	1.6	42.3	1.3	47.9
BE9R	14.8	0.5	15.4	0.3	2.8	0.1	7.9	2.4	48.5	1.5	56.4
BE9S	16.4	0.4	8.5	0.2	1.4	0.1	5.6	1.6	42.3	1.3	47.9

There is an obvious problem with the above table. It has been assumed that the total A-level points score is based on the current points system. Obviously similar issues as were raised for aggregating GCSE results could be considered. There is need for research because of recent proposals to change the A-level points system.

Using aggregate measures to predict results for individual A-level subjects

In the following sub-section, individual A-level subjects have been considered. English and mathematics were selected for analysis in this section. A more extensive set of analyses including other subjects is available in Bell (2000). The analyses described in this section used five samples, one for each subject, of all candidates from a random sample of 250 A-level centres.

It is inappropriate to fit an ordinary multilevel model to an ordinal outcome such as an A-level grade. Although it is possible to model A-level grades with proportional odds models (Fielding, 1999), this requires strict assumptions that do not seem to hold in practice. For this reason, the binary response variables will be analysed using multilevel logistic regression. For the purposes of this paper, two grade boundaries, A/B, and E/U, have been considered, i.e., the probability of being 'at grade A' and the probability of getting 'at least grade E'. The issue of the type of model for examination grades is considered in greater detail in Bell and Dexter (2000a, 2000b).

The following multilevel model was fitted.

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 meas_{ij} + \gamma_{0i} + \varepsilon_{ij}$$

where p_{ij} is the probability of candidate *j* from centre *i* obtaining at least grade *x*,

β_0 is a constant,

β_1 is a slope parameter,

$meas_{ij}$ is an aggregate measure of GCSE performance,

γ_{0i} is a random centre level error term,

ε_{ij} is a random candidate level error term.

The multilevel model described above was fitted for the two dependent variables ('at grade A' and 'at least grade E') for the null or empty model and for each of the fifteen measures and for both subjects. These two boundaries were chosen because they are boundaries determined by judgement in the grade awarding process. The full results for the analyses are given in the Appendix. In addition to the estimation of the model parameters by MLwin the explained variance and the unexplained variance at the centre and candidate level were calculated using the procedures described in Snijders and Bosker (1999).

The explained and unexplained variance statistics for both subjects at grade A and at least grade E are presented in Table 6. The measure that explains the most variation has been highlighted in bold. For example, BE5R explains 57% of the variation for mathematics. As a measure BE5R favours candidates who have a consistent performance on five subjects. For grade A the proportion of centre level variation is relatively small for both subjects. A possible explanation for this is that these high performing candidates are taking most of the responsibility for their learning. (This a characteristic of higher levels of education. For a discussion of levels, see Bell and Greatorex, 2000.) The differences in the amount of centre level variation between Mathematics and English, for example, could reflect the nature of assessment in these subjects. For mathematics, the marking is objective and it is clear what a candidate has to do to achieve well and for English, the marking is more subjective and a candidate would need more guidance to produce good answers. It should also be noted that the amount of unexplained centre level variation is highest for TOT, TOTR and TOTS. These measures are influenced by centres' policies on the number of GCSEs they allow Y11 pupils to sit.

Table 6: Explained variation for the logistic models for the dependent variables 'at grade A' and 'at least grade E'

	Grade A						At least grade E					
	Mathematics			English			Mathematics			English		
Measure	EXP	UC	UP	EXP	UC	UP	EXP	UC	UP	EXP	UC	UP
None	0.00	0.15	0.85	0.00	0.19	0.81	0.00	0.23	0.77	0.00	0.18	0.82
MN	0.47	0.05	0.48	0.39	0.09	0.52	0.28	0.14	0.58	0.31	0.10	0.58
MNR	0.49	0.04	0.47	0.43	0.08	0.49	0.24	0.15	0.61	0.29	0.10	0.61
MNS	0.45	0.05	0.50	0.35	0.10	0.55	**0.33**	**0.15**	**0.53**	**0.38**	**0.09**	**0.53**
TOT	0.32	0.10	0.58	0.29	0.11	0.61	0.12	0.22	0.66	0.16	0.13	0.71
TOTR	0.18	0.14	0.68	0.16	0.14	0.70	0.04	0.23	0.72	0.05	0.16	0.80
TOTS	0.39	0.07	0.54	0.32	0.10	0.58	0.24	0.18	0.58	0.32	0.10	0.58
MN9	0.46	0.04	0.50	0.41	0.08	0.51	0.14	0.20	0.66	0.22	0.11	0.67
MN9R	0.50	0.04	0.47	0.47	0.06	0.46	0.06	0.22	0.72	0.11	0.13	0.76
MN9S	0.44	0.05	0.51	0.35	0.09	0.56	0.26	0.16	0.58	0.33	0.10	0.57
BE5	0.51	0.05	0.45	0.41	0.09	0.50	0.15	0.19	0.66	0.20	0.12	0.68
BE5R	**0.57**	**0.03**	**0.41**	**0.52**	**0.07**	**0.42**	0.06	0.22	0.72	0.11	0.14	0.75
BE5S	0.47	0.05	0.48	0.35	0.10	0.55	0.24	0.16	0.59	0.29	0.11	0.60
BE9	0.46	0.04	0.49	0.41	0.08	0.51	0.13	0.21	0.66	0.20	0.12	0.68
BE9R	0.51	0.03	0.46	0.50	0.06	0.44	0.05	0.23	0.72	0.06	0.14	0.79
BE9S	0.44	0.05	0.51	0.35	0.09	0.56	0.25	0.17	0.58	0.32	0.10	0.58

(EXP – explained variance, UC – Unexplained variance at the centre level, UP – unexplained variance at the candidate/pupil level)

These results for 'at least grade E' are different from those for grade A. Firstly, the amount of variation explained is much smaller. Secondly, the best aggregate measure is MNS rather than BE5 and, thirdly, the amount of unexplained centre level variation is much greater. The results for grade E mean that candidates who do badly on most of their GCSE subjects but very well on a few have a greater probability of success at A-level than candidates who have consistent mediocre grades. This is plausible because candidates have specialised at A-level and will have opted to study subjects that are closely related to those GCSE subjects

in which they did well. The greater level of centre variation would suggest the centres are more likely to influence the performance of weaker candidates

Discussion

In this paper it has been demonstrated that the scoring system chosen to aggregate GCSE results can have a significant impact on the amount of variation in A-level results that can be explained. It is also possible to devise a number of different measures of aggregate GCSE performance. It has been demonstrated that these measures have different characteristics and can be influenced by school policies on examination entrance. The choice of measure could change the rank order of centres in a performance table.

From preliminary analyses described in this paper, it would seem that the differences between most of the measures of aggregate GCSE performance in predicting the total A-level score are not great. However, this was not the main focus of the paper. There is a need to investigate the properties of these measures with different aggregate measures of A-level performance, large samples and more complex models.

The main focus of this paper (and the main emphasis for awarding bodies) is the prediction of performance in individual A-level subjects. The results of the multilevel analyses were particularly interesting because different aggregate measures were found to be appropriate for predicting high A-level grades compared with low A-level grades. The best measure for predicting good A-level grades was BE5R, which is calculated by taking the square root of the GCSE points score and adding up the transformed points for the best five results. This measure favours candidates who perform consistently on a limited range of, presumably relevant, subjects. This could also explain some of the reduction of the centre level variation. School and national policies (e.g., the National Curriculum which only applies to state-maintained schools) could result in candidates being required to enter subjects in which they cannot do well. For example, some centres might only offer double award GCSE science while others might offer a choice between single and double award. Candidates who disliked science could have lower aggregate measures of GCSE performance for the first set of centres compared with the second set for those measures that include all GCSE results. This could result in spurious conclusions about the differential centre effectiveness of A-level teaching.

For lower levels of A-level performance, the best measure was MNS. This is the arithmetic mean of the squares of GCSE scores. This measure favours erratic candidates. The fact that an aggregation that rewards erratic performances has the most explanatory power is probably related to the fact that candidates specialise at A-level. Candidates choose A-levels in subjects that are related to their best GCSE results (Bell, 2000b). This means that an erratic candidate would tend to have higher GCSE grades in the subjects they specialise in at A-level than might a steady candidate (e.g., AAACCCEEE compared with CCCCCCCCC where underlining denotes the subject for A-level). These results apply to a wider range of subjects than present here (see Bell, 2000a).

The specific conclusion is that the choice of aggregate measure is important and that the assumption that the mean GCSE score is the best predictor is unjustified. There is a need for further research with these measures, particularly, with the more complex models used to investigate substantive issues such as a school effectiveness. More generally, this paper illustrates the issues involved in aggregating scores to produce simple summary statistics.

References

Bell, J.F. (1998). *Patterns of subject uptake and examination entry 1984-1997*. Paper presented at the British Educational Research Association Annual Conference. Belfast: Queen's University (Available from Education-line at http://www.leeds.ac.uk/educol/index.htm).

Bell, J.F. (In press) Patterns of subject uptake and examination entry 1984-1997. *Educational Studies*,

Bell, J.F. (2000) *Methods of aggregating GCSE results to predict A-level performance*. Paper presented at the British Educational Research Association Annual Conference. Cardiff: University of Wales (Available from Education-line at http://www.leeds.ac.uk/educol/index.htm).

Bell, J.F., & Dexter, T. (2000a). *Using multilevel models to assess the comparability of examinations*. Paper to be presented at the Fifth International Conference on Social Science Methodology of the Research Committee on Logic and Methodology (RC33) of the International Sociological Association (ISA). Cologne: University of Cologne. (Available from Education-line at http://www.leeds.ac.uk/educol.htm).

Bell, J.F., and Dexter, T. (2000b) Using ordinal multilevel models to assess the comparability of examinations. *Multilevel modelling newsletter*, December, 12, 2, 4-9. (Available online at http://www.ioe.ac.uk/multilevel/new12-2.pdf)

Bell, J.F., & Greatorex, J. (in prep.). *A review of research in levels, profiles and comparability*. London: QCA (available online at http://www.qca.org.uk).

Department for Education (1995) *The development of a National Framework for estimating value added at GCSE A/AS level. Technical Annex to GCSE to GCE A/AS value added: briefing for schools and colleges.* London, Department of Education).

Fielding, A. (1999). Why arbitrary points scores? Ordered categories in models of educational progress. *Journal of the Royal Statistical Society, Series A - Statistics in Society*, 163, 3, 303-328.

Gray, J., Goldstein, H., & Jesson, D. (1996). Changes and improvements in schools' effectiveness: trends over five years. *Research Papers in Education*, 11, 1, 35-51.

Gray, J., Hopkins, D., Reynolds, D., Wilcox, B., Farrell, S., & Jesson, D. (1999*). Improving Schools. Performance and Potential*. Buckingham: Open University Press.

Jones, B. E. (1997). Comparing Examination Standards: is a purely statistical approach adequate? *Assessment in Education*, 4, 2, 249-262.

O'Donoghue, C., Thomas, S., Goldstein, H., and Knight, T. (1997) 1996 DfEE study of value added for 16-18 year olds in England. *DfEE research series.* (London, Department for Education and Employment).

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Wood, R. (1991). *Assessment and Testing: A Survey of Research*. Cambridge: Cambridge University Press.

Yang, M., and Woodhouse, G. (2001) Progress from GCSE to A and AS level: institutional and gender differences, and trends over time. *British Educational Research Journal*, 27, 3, 245-267.

Appendix: Results of logistic regression analyses

Note that for each sample the standardisation was carried out using the mean and standard deviation of the samples. This means that the constant parameter can be interpreted as the logit of the probability of obtaining at least a grade *x* for an average pupil from an average centre. The means and variances vary from sample to sample because of sampling error and the fact that the entries for subjects differ. Some of the analyses were carried out with MLwin and other analyses were carried out using MIXOR because of estimation problems with MLwin. MLwin had a tendency to converge to a solution with zero centre variance for some data sets. The centre level standard error for the analyses carried out using MIXOR is for the standard deviation at the centre level rather than the variance.

Table A1: Multilevel model results for mathematics with dependent variable 'at grade A' (MIXOR)

Measure	Const.	s.e.	slope	s.e.	centre	s.e.
None	-1.33	0.07			0.58	0.07
MN	-1.77	0.07	1.80	0.05	0.31	0.08
MNR	-1.78	0.07	1.85	0.05	0.29	0.08
MNS	-1.74	0.07	1.73	0.05	0.33	0.08
TOT	-1.57	0.07	1.35	0.04	0.56	0.08
TOTR	-1.46	0.08	0.94	0.03	0.66	0.08
TOTS	-1.63	0.07	1.54	0.05	0.43	0.09
MN9	-1.68	0.06	1.74	0.04	0.28	0.08
MN9R	-1.62	0.06	1.86	0.03	0.25	0.08
MN9S	-1.69	0.06	1.67	0.05	0.33	0.08
BE5	-1.79	0.06	1.93	0.04	0.34	0.08
BE5R*	**-1.72**	0.06	**2.14**	**0.03**	**0.23**	**0.08**
BE5S	-1.80	0.07	1.80	0.05	0.32	0.08
BE9	-1.69	0.06	1.76	0.04	0.27	0.08
BE9R	-1.63	0.06	1.92	0.03	0.24	0.08
BE9S	-1.70	0.07	1.68	0.05	0.32	0.08

Table A2: Multilevel model results for mathematics with dependent variable 'at least grade E' (MLwin)

Measure	const	s.e.	slope	s.e.	centre	s.e.
None	1.75	0.08			0.98	0.15
MN	2.42	0.10	1.26	0.06	0.82	0.15
MNR	2.31	0.09	1.13	0.06	0.81	0.14
MNS	**2.54**	**0.10**	**1.43**	**0.07**	**0.91**	**0.14**
TOT	2.02	0.09	0.78	0.06	1.08	0.16
TOTR	1.85	0.09	0.45	0.05	1.06	0.16
TOTS	2.31	0.10	1.18	0.07	1.01	0.16
MN9	2.11	0.09	0.84	0.06	0.99	0.16
MN9R	1.91	0.09	0.51	0.05	1.03	0.16
MN9S	2.37	0.10	1.21	0.07	0.91	0.15
BE5	2.10	0.09	0.87	0.06	0.95	0.15
BE5R	1.89	0.09	0.52	0.06	1.00	0.15
BE5S	2.31	0.10	1.16	0.06	0.91	0.15
BE9	2.09	0.09	0.82	0.06	1.03	0.16
BE9R	1.90	0.10	0.49	0.05	1.05	0.16
BE9S	2.34	0.10	1.18	0.07	0.97	0.16

Table A3: Multilevel model results for English with dependent variable 'at grade A' (MLwin)

Measure	Const.	s.e.	slope	s.e.	centre	s.e.
None	-2.04	0.08			0.76	0.13
MN	-2.76	0.10	1.58	0.07	0.54	0.12
MNR	-2.79	0.10	1.70	0.08	0.51	0.11
MNS	-2.68	0.09	1.45	0.06	0.57	0.12
TOT	-2.49	0.08	1.25	0.06	0.57	0.11
TOTR	-2.26	0.08	0.86	0.06	0.64	0.12
TOTS	-2.51	0.09	1.34	0.06	0.54	0.12
MN9	-2.74	0.09	1.61	0.08	0.51	0.11
MN9R	-2.75	0.09	1.83	0.10	0.46	0.10
MN9S	-2.67	0.09	1.44	0.06	0.55	0.12
BE5	-2.76	0.10	1.64	0.08	0.56	0.12
BE5R	**-2.81**	**0.10**	**2.02**	**0.09**	**0.54**	**0.12**
BE5S	-2.69	0.09	1.45	0.06	0.57	0.12
BE9	-2.76	0.09	1.62	0.08	0.50	0.11
BE9R	-2.76	0.09	1.93	0.10	0.45	0.10
BE9S	-2.67	0.09	1.44	0.06	0.53	0.12

Table A4: Multilevel model results for English with dependent variable 'at least grade E' (MLwin)

Measure	const	s.e.	slope	s.e.	centre	s.e.
None	2.08	0.08			0.70	0.13
MN	2.77	0.10	1.33	0.07	0.57	0.12
MNR	2.61	0.09	1.25	0.06	0.56	0.12
MNS	**2.92**	**0.10**	**1.55**	**0.08**	**0.57**	**0.12**
TOT	2.40	0.08	0.85	0.06	0.61	0.12
TOTR	2.19	0.08	0.44	0.04	0.64	0.12
TOTS	2.72	0.09	1.34	0.08	0.58	0.12
MN9	2.53	0.09	1.03	0.06	0.55	0.12
MN9R	2.30	0.08	0.68	0.06	0.56	0.12
MN9S	2.79	0.08	1.39	0.08	0.56	0.12
BE5	2.48	0.09	0.98	0.07	0.59	0.12
BE5R	2.64	0.08	0.69	0.07	0.60	0.12
BE5S	2.68	0.09	1.25	0.07	0.59	0.12
BE9	2.49	0.09	0.98	0.06	0.58	0.12
BE9R	2.26	0.08	0.51	0.05	0.59	0.12
BE9S	2.76	0.10	1.36	0.08	0.58	0.12