RELATIVE RATING – A MEANINGLESS MEASURE

R.A.SPARKES

SYNOPSIS

A common indicator used to determine the 'performance' of a school subject department is its *Relative Rating*, which compares the mean grade that its candidates achieve in the public examination of that subject with the mean grade that they achieve in all the other subjects they are presented for at the same time. Since it is the *same* candidates, it is assumed that any difference between these mean grades is due to the effectiveness of their subject departments - a positive rating identifying a department that is 'above average'. Our study shows that this argument is flawed, firstly because candidates perform so inconsistently across their different subjects, that only a small proportion of the 'rating' can be attributed to departmental effectiveness and secondly because a department's *Relative Rating* is only reliable if its candidates are exactly representative of *all* candidates nationally, which is unlikely to be true. It is therefore concluded that this indicator is completely useless for its intended purpose.

THE ATTRACTION OF THE *RELATIVE RATING*

School performance indicators have encountered a great deal of criticism, much of which is dismissed as professional contempt for public accountability. This paper takes the view that it is the *indicators* which are wrong, not the principle of evaluating teachers' performance. Indicators based upon the proportion of a year group attaining a particular grade in public examinations (e.g. the National Comparison Factors, or NCFs) – commonly called 'raw results' – are published nationally, but the general academic view is that, because they take no account of the abilities, aspirations and motivation of candidates, they are not a valid measure of school or departmental effectiveness (Nuttall, Blackhouse and Willmott, 1989; Smith and Tomlinson, 1989; Kendall, 1995; Fitz-Gibbon, 1996; Gray, 1996; Murphy, 1997; Gibson and Asthana, 1998).

An attempt has been made to allow for differences between candidates by introducing *Value Added* methods (Fitz-Gibbon, 1992; Gray, Allnutt, Gardner, Blackham and Frost, 1995; Thomas and Mortimore, 1996; Woodhouse and Goldstein, 1996; O'Donaghue, Thomas, Goldstein and Knight, 1997; SOEID, 1998), where 'prior performance' is used as a reference and which thus concentrates attention on the '*progress*' made by each pupil. For example, in Scottish secondary schools a candidate's 'progress' between Standard grade and Higher grade is determined by the amount by which his or her Higher grade result in a particular subject is better or worse than the average result obtained by similar candidates. In this context, candidates are said to be 'similar' if they obtained the same mean for all their Standard grade results in the previous year - a figure which is called their GPA, or grade point average. (Why two candidates, who study different subjects and different numbers of subjects, can ever be described as 'similar' merely on the basis of their GPA is an unsolved mystery!) The average 'progress' made by a school's candidates in each subject is called the *Value Added* indicator of that subject department. This is regarded as a fairer measure of school effectiveness than 'raw results' (Thomas, et al. 1998), however, it has been shown (Tymms and Fitz-Gibbon, 1990; Thomas and Mortimore, 1996; Goldstein and Spiegelhalter, 1996; Gray, 1996; Woodhouse and Goldstein, 1996; Coe and Fitz-Gibbon, 1998; Sparkes, 1999) that the *Value Added*

indicator fluctuates widely from year to year, because the number of candidates in a department is rarely sufficient to even out the random influences that affect each individual pupil. This means that a subject department needs to be consistently and significantly above (or below) average before it can be regarded as worthy of commendation (or concern) and this is only true for a small number of departments. Furthermore, there is such a high correlation between 'raw results' and *Value Added* indicators that the latter are clearly still dependent on the social context of the school's pupils and are therefore poor indicators of teacher 'performance'.

One performance indicator that appears to avoid this problem is the *Relative Rating* of a school subject department, where the examination results in different subjects are compared *within a single school* (SOEID, 1991, 1998). Because it the *same* candidates who are being compared, their social context, abilities, aspirations and motivation, etc. are assumed to be the same for all departments and thus should not confound the measurement. Hence it ought to be a better indicator of the effectiveness of a subject department. The interpretation of the Scottish Executive Education Department (SEED) is that school departments with positive ratings are performing 'above average' whereas negative ratings "may raise issues concerning the quality of courses, approaches to learning and teaching or for support for pupils …" (SOEID, 1998, p. 7). Unfortunately, this is an illusion.

DETERMINING A DEPARTMENT'S *SCHOOL RATING*

In Scotland, Year 11 pupils (15 to 16 year olds) sit the Standard grade examination in a range of subjects, similar to the GCSE in England and Wales. Awards are made by the Scottish Qualifications Agency (SQA) in seven grades. The average grade obtained by a candidate in all the subjects taken at one sitting is his or her GPA (grade point average). Candidates are normally presented for the Higher grade examination one year later, which is awarded at grades A, B, C, D and No Award. In practice, 'Highers' are actually assessed on the numerical scale 1 to 14 (called *range-scores*) and these are converted to letter grades according to the following code: 1-5 = grade A, 6-7 = B, 8-9 = C, 10-11 = D and 12 to 14 = No Award. Because range-scores represent a finer grading, they are normally preferred for statistical analysis and *Relative Rating* and *Value Added* indicators are actually calculated in range-scores. To allow them to be compared with Standard grade, the SQA then halve their figures to make them equivalent to grades. This dubious practice (range-scores are **not** twice the size of their corresponding letter grades) was circumvented in our investigation by doubling the SQA figures to convert them back to range-scores.

The Higher examination results of all the candidates in a particular subject department, who were also presented for at least one other subject, are averaged to obtain a mean range-score for that subject. The mean range-score that these same candidates obtained in all their other subjects is also determined and the difference between them produces the *School Rating* for that subject. Taking English as an example, for all candidates in a school who were presented for English:

> *School Rating* for English = mean range-score in the *other* subjects – mean range-score in English

Candidates in an English department with a positive *School Rating* have thus achieved better results in English (lower range-scores) than they have in their other subjects, hence the conclusion that this English department could be congratulated on its 'above average' performance.

THE *NATIONAL RATING* OF A SCHOOL SUBJECT

Unfortunately, this interpretation is invalidated by the discovery that nearly all

English departments have a positive *School Rating* and that nearly all Mathematics departments have a negative one. Since it is undiplomatic to infer from this that Mathematics teachers generally are 'below average', an alternative explanation has been developed. It is assumed that the 'poorer' performance of Mathematics departments is because this subject is 'more difficult'. The notion of 'subject difficulty' has been around for some time (Nuttall, 1974; Bardell, *et al*, 1978; Forrest and Vickerman, 1982; Tymms and Fitz-Gibbon, 1991; Fitz-Gibbon and Vincent, 1994) and there has recently been a debate on this issue between Fitz-Gibbon and Vincent (1997) and Goldstein and Cresswell (1996). For the purpose of comparing subject departments, this idea is very appealing, it is simply assumed that every subject ought to have the same mean grade (after differences in the abilities of each subject's candidates have been allowed for). A 'correction factor' for each subject is therefore calculated, which, when added to each candidates' score, produces just this result. A technique for calculating these 'correction factors' has been developed by Lawler, contained in an appendix to Kelly (1976) and it has been the annual practice of the SQA for many years to publish these Kelly 'correction factors' for each subject - calling them *National Ratings*. The *National Rating* for Mathematics is negative, so it is termed a 'more difficult' subject, whereas the *National Rating* for English is positive – it is 'easier'. However, whether this is because Mathematics is inherently more difficult to learn or because it is less interesting or motivating than English or because it is less well taught or because its examination is more severely graded is not known.

THE *RELATIVE RATING* OF A SCHOOL'S SUBJECT DEPARTMENT

The *National Rating* for each subject is subtracted from a department's *School Rating* to produce its *Relative Rating* – a figure that may be used to compare that department with others in the same school. An example should clarify this process. Suppose that the *National Rating* for a particular subject is –0.5 (meaning that it is half a range-score 'more difficult' than the average of the other subjects) and one particular department's *School Rating* for this subject is +1.0 (meaning that its candidates performed one range-score better in this subject than the average of their other subjects), then the *Relative Rating* for the department is +1.5 (this department achieved one-and-a-half range-scores better than the other departments in the same school).

This incorporation of a subject's *National Rating* with the subject department's *School Rating* effectively removes a powerful argument against the use of the latter. Unfortunately, if Mathematics really is poorly taught compared with other subjects, this inclusion of its *National Rating* will effectively hide the fact. This alone should be enough to cast doubt upon the whole process, but there is more!

INCONSISTENCY IN THE *RELATIVE RATING*

In both the *Value Added* and the *Relative Rating* indicators, some of the random factors that influence each individual candidate such as the uncertainty of grading his or her examination scripts (Mortimore and Mortimore, 1984), will tend to cancel out when results are averaged. They are unlikely to cancel out completely, though, so that performance indicators will show some uncertainty. If individual results are influenced in the *same* direction (for example, by an 'effective' department), then the indicator might be expected to reveal this 'effectiveness'. Unfortunately, the uncertainty generally swamps the influence of the school or department. How does this uncertainty arise?

If fifty coins are spun at the same time, it is unlikely that they will land with *exactly* 25 'heads' and 25 'tails'. It would be ridiculous to identify the coin spinner as 'below average' if he only scored 20 'heads', yet this is effectively what the

SEED interpretation of the *Relative Rating* does. Probability theory enables us to determine that our coin-thrower could expect to get somewhere between 18 and 32 'heads' on 95% of the occasions that he throws the coins, that is, 19 times out of every 20. This range, denoted as 25 ± 7, is called the *95% confidence interval* for this experimental result.

It is possible to determine a similar 95% confidence interval for any particular *Relative Rating*. It depends upon the number of candidates in the department – the smaller the number, the wider the interval. For an average-sized department, the figure is about ± 0.7 range-scores (0.35 grades), so that an 'average performance' department could expect to see its *Relative Rating* fluctuate between –0.35 and +0.35 grades from year to year. Although the SEED recognises this principle, it does not apply it consistently. It notes, for example, that the *Relative Rating* for a small department is too unreliable, so that it does not publish ratings for departments with fewer than 10 candidates. If a department's *Relative Rating* is outwith the 95% confidence interval for its number of candidates, the figure is asterisked (and described as '*notably*' above or below average). However, the level required for a figure to be asterisked is only set at 95% , which gives a 1 in 20 chance that a department could fluctuate above or below this 'notable' level by random chance. With over 400 schools, each with at least 20 departments, this could result nationally in over 400 subject departments each year being unfairly designated. It is therefore essential to consider more than one year's results, which is why the SEED produces figures for three years at a time. Our objection, though, is that the SEED still publishes the figures even when these are *within* the 95% confidence interval – it ought at least to indicate that such ratings are indistinguishable from zero and that such departments should, therefore, be designated as 'average'.

THE PRESENT INVESTIGATION

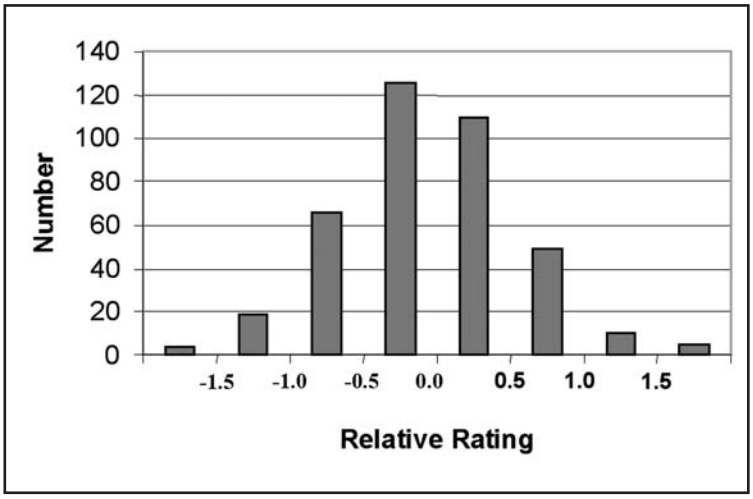
It was decided to test the validity of this interpretation by working with actual examination results.

In 1997, around 32,000 candidates were presented for the Higher grade in over 40 subjects. A full set of these results was obtained from the SQA in order to determine the fluctuation in a subject department's *Relative Rating* that might be expected from year to year. English was the subject chosen for the investigation, on the grounds that this is the most popular subject and could thus be expected to have the smallest uncertainty. It was noted that the difference between the range-scores that individual candidates attained in English and their mean range-scores in their other subjects varied between –5.0 and +6.0, clearly demonstrating that candidates do not perform consistently across subjects. *Relative Ratings* for English were calculated for all schools with more than 10 candidates and their distribution is shown in Figure 1.

Because *Relative Ratings* are based on an average of individual candidates' results, the aforementioned inconsistency is considerably reduced. If it is assumed that schools have no differential influence on their candidates' achievements, the 95% confidence interval for this distribution should theoretically be 0.65. In practice, it was found to be 1.25, much larger, because departments **do** have differential influence on their candidates' results. The difficulty is in discriminating between the proportion of this influence that can be attributed to their teachers and the proportion that can be attributed to external social factors over which they have no control.

When the researcher investigated the *Value Added* indicator (Sparkes, 1999), he was able to create a simulation of the results that could be expected from a set of schools where the 'candidates' and their 'teachers' remained unchanged for 25 years. This enabled a calculation of the annual fluctuation in a subject department's *Value Added* indicator to be made. It was decided to carry out the same simulation to determine the annual fluctuation that might occur in a department's *Relative Rating,*

Figure 1: The Distribution of School Relative Ratings for English



concentrating on Mathematics and English and effectively measuring the *difference* between the *Relative Ratings* of these two subjects.

The average school has 29 candidates presented for both Mathematics and English, and a typical school of this size was chosen as representative. Simulated results were produced for each candidate in this school by using the following formulae for the range-scores obtained by each candidate.

$$\text{English}_i = 2.33 + 2.70 * \text{GPA}_i + \varepsilon_i$$
$$\text{Mathematics}_i = 2.36 + 3.22 * \text{GPA}_i + \varepsilon_i$$

where GPA_i is the GPA of the ith candidate. The same set of 29 GPAs was used each time to ensure that it was the *same* candidates that were being presented for the examination in each of the 25 simulated occasions. The first parts of these formulae are the regression equations published by the SQA (and used for calculating *Value Added* indicators). The regression equations predict that a candidate with the average GPA of 2.0 would achieve a range-score of 7.7 in English and 8.8 in Mathematics, giving English a 'correction factor' of 1.1 range-scores when compared with Mathematics (the difference between the SQA *National Ratings* of these subjects is 0.51 grades, or 1.02 range-scores).

The *actual* results of candidates are approximately normally distributed. In the simulated results this is reproduced by the introduction of the ε_i term, which is a normally distributed random number with a mean of 0.00 and a standard deviation of 1.65. This term represents the numerous factors that influence each candidate's individual results, such as his or her motivation, preparedness, emotional state, etc., which are so varied and so numerous that they can be considered random. When the model is run, it produces simulated range-scores for each candidate in both subjects, and the difference between the two mean range-scores should correspond to the difference between the SQA *National Ratings* for these subjects.

The model was run 25 times for the same 29 candidates and the difference between the English and Mathematics means was calculated each time. It varied between 0.2 and 1.7, with a mean of 1.0. If the latter is taken as the *School Rating* difference between these subjects, when the *National Rating* difference of

1.02 is subtracted, the school's English-Mathematics *Relative Rating* difference becomes zero. In this school, therefore, both departments perform at the same level, on average. For our purposes, the important point was the *variation* in this difference – it had a 95% confidence interval of ± 0.74. This can be taken as the annual fluctuation to be expected in the difference in ratings between English and Mathematics. The simulation therefore reinforces the points made previously – the *Relative Rating* of any subject department can be expected to fluctuate because of the many individual factors over which the school has no control and candidates' individual inconsistencies do not cancel out when their results are averaged. It is therefore quite unfair to criticise a subject department for a single 'below average' result. Only if a department's *Relative Rating* is 'notably' and *consistently* above (or below) average would it need to be elated (or concerned). Even then, the determination does not indicate whether this occurs because these departments do a relatively better (or worse) job of preparing their candidates for the examination, or whether it is because of some other factors over which the departments have no control. (It may be, for example, that candidates in some areas see more chance of employment with certain subjects than with others, so they are more motivated to study them.)

PROBLEMS WITH THE *NATIONAL RATING*

To allow for possible differences in the 'difficulty' levels of different subjects, the *Relative Rating* takes into account the *National Rating* of each subject. Unfortunately, this is a single figure for each subject, yet a subject's 'difficulty' is bound to depend upon the individual circumstances of each candidate. Again, the assumption is made that such individual differences will cancel out when results are averaged, but the previous discussion should warn us of the dangers of doing this. The number of candidates in each department is far too small for it to happen effectively. Let's first see how the 'difficulty' of a subject varies between different groups of candidates.

Ability differences

At the Higher grade, the calculation of a subject department's *Relative Rating* from its *School Rating* assumes that the subject's *National Rating* is a **single** figure, i.e. the subject has the same relative 'difficulty' for all candidates. When the SQA publishes *National Ratings* for Standard grade subjects, it produces both 'upper' and 'lower' figures. The former is obtained by including only those candidates who obtained at least three Standard grades at a grade of 1, 2 or 3 (i.e. the more academically 'able' candidates), the 'lower' *National Rating* is obtained from the others. The two figures differ by a substantial amount, for example, the 'upper' rating for English is 0.20 and the 'lower' is 0.70, with an overall value of 0.36. Hence, an English department with a *School Rating* of +0.28 would find this figure reduced to -0.08 if the overall *National Rating* is subtracted (the school is *below* average). However, if the school were in an affluent area, with high achieving pupils, most of whom achieved three or more 'good' Standard grades, the 'upper' rating (0.20) would have been subtracted instead, so the actual school's *Relative Rating* would have been +0.08 (it is now *above* average). On the other hand, if the school had been in a socially and economically deprived area, the 'lower' *National Rating* (0.70) would probably have been used, resulting in a *Relative Rating* of –0.42 ('*notably*' below average).

Because of this, the SEED provides *Relative Ratings* for Standard grade results based upon these 'upper' and 'lower' ratings. Why, though, are only **two** *National Ratings* used? Candidates have a much wider range of 'abilities' than this simple two-way categorisation can accommodate.

This phenomenon was first investigated for the Higher grade examination without

analysing actual Higher grade results, since the SQA publishes regression equations to enable schools to determine their own performance indicators. It was possible to use these to determine the *expected* grades in each Higher subject for candidates with GPAs of 1.0, 2.0 and 3.0. Table 1 shows the difference in the predicted range-scores for each 'ability' level, along with their SQA *National Ratings* (multiplied by two to convert them to range-scores). The table also shows the proportion of schools which have mean GPAs in each of these three categories.

Table 1: Comparison of the Value Added predictions for three levels of 'ability' with the SQA National Ratings for English and Mathematics

Subject	GPA			SQA *National Rating*
	1.00	2.00	3.00	
Proportion of schools	2%	85%	13%	
English	5.03	7.73	10.43	0.26
Mggathematics	5.58	8.80	12.02	-0.76
difference (Eng. – Maths.)	-0.55	-1.07	-1.59	-1.02

It can be seen that the difference between the predicted *National Ratings* for English and Mathematics varies according to candidates' 'ability'. Both subjects have a different level of 'difficulty' for candidates of different 'ability', so that a **single** SQA *National Rating* would be inaccurate for 'high achieving' and 'low achieving' schools by about 0.5 of a range-score.

ii) Gender differences

This 'ability' difference in a subject's 'difficulty' was further investigated by categorising candidates according to sex as well as 'ability'. The Higher grade datafile was divided into males and females and then split into five 'ability' groups. The equivalent of *School Ratings* for each subject (now called 'Group Ratings') were calculated separately for each category. They measure the difference between the mean achievement of each group in the chosen subject and their mean achievement in all their other subjects, exactly as with the *School Rating*. The results for Mathematics and English are shown in Table 2.

The 'Overall' figures are the weighted mean of the separate ratings for boys and girls (different proportions of each sex and each ability group were presented for each subject) and these may be compared with the SQA *National Ratings* (doubled to convert them to range-scores). Although the *numbers* of females in each ability group were different from the numbers of males (the proportion of females in the 'high ability' group exceeded 60%), the *mean* GPAs of the sexes in each ability group were identical. The figures therefore show that the differences between the 'Group Ratings' for boys and girls are not just due to differences in ability, there is a sex difference too. This clearly has implications for single sex schools – their *Relative Ratings* will be inaccurate by anything up to one range-score when the single *National Rating* is used to correct for 'subject difficulties'. Incidently, the figures also confirm the general belief that boys find English 'more difficult' than girls of the same 'ability', but find Mathematics 'easier'.

Table 2: 'Group Ratings' in English and Mathematics for each sex at each 'ability' level

GPA categories	'Group Ratings' for English			'Group Ratings' for Maths.		
	Male	*Female*	*Combined*	*Male*	*Female*	*Combined*
Less than 1.50	-0.25	-0.08	-0.14	-0.32	-0.64	-0.53
1.50 to 1.99	0.15	0.14	0.17	-0.46	-0.79	-0.62
2.00 to 2.49	0.19	0.40	0.32	-0.72	-1.14	-0.91
2.50 to 2.99	0.05	0.47	0.29	-0.84	-1.29	-1.05
Greater than or equal to 3.0	0.21	0.60	0.44	-1.16	-1.72	-1.34
SQA *National Ratings*	0.26			-0.76		

iii) 'Interest' differences

A third difference was found by categorising the candidates according to their subject specialisms. Two subgroups were extracted from the datafile of all those who had been presented for at least four subjects. Candidates who had, in addition to English and Mathematics, also been presented for at least two of Biology, Chemistry, Computing Studies, Geography, Human Biology, Physics and Technological Studies were designated 'science' and those who had also been presented for at least two of Art & Design, French, German, History, Latin, Modern Studies, Religious Studies and Spanish were called 'arts'. The mean GPAs of the candidates in each category were almost identical (1.51 and 1.49), although there were 3626 'science' candidates compared with 1463 'arts' candidates. Their mean range scores over all their Higher grade subjects were also very similar, furthermore, the proportion of each sex was about the same for both groups. For each group, 'ratings' for English and Mathematics were determined as before. The figures generally showed that 'arts' candidates find Mathematics 'more difficult' and English 'easier' than candidates in general, whereas 'science' candidates find the reverse (Table 3). This phenomenon is not explained by a difference in 'ability' nor gender, since the two groups are alike in this respect, the only difference is 'interest', illustrating the unsurprising point that candidates tend to choose their Higher subjects from those in which they personally achieve better results.

Table 3: 'Group Ratings' for English and Mathematics for 'arts' and 'science' candidates.

	'Group Ratings'			SQA *National Rating*
	Arts	Science	Difference (Arts–Science)	
Mathematics	-1.28	-0.45	-0.83	-0.76
English	0.94	0.19	0.75	0.26

There is evidence from these figures that, on average, Mathematics is overall 'more difficult' than English; even the 'science' candidates found English more than half

a range-score 'easier' than Mathematics. These results correspond with the findings of Forrest and Vickerman (1982) and Fitz-Gibbon and Vincent (1997) that A level Mathematics is more 'difficult' than English (although, in the case of A levels, the difference was greater).

Two extreme groups were also compared. All candidates who had attained grade A in Higher Physics or Chemistry – the physical sciences group – were selected and the 'Group Ratings' for all their subjects were calculated as before. The same process was repeated for candidates who had attained grade A in Higher French, German or Spanish – the modern foreign languages group. The mean GPAs of the two groups were again very similar (physical sciences group = 1.31, modern foreign languages group = 1.21) so any differences between the two groups are unlikely to be the effect of different 'abilities', although the anticipated gender imbalance between the two groups was clearly observed. Table 4 shows the 'Group Ratings' for English and Mathematics along with their SQA *National Ratings*. The physical science candidates actually find English 'more difficult' than Mathematics and the modern foreign languages candidates find Mathematics their 'most difficult' subject.

Table 4: 'Group Ratings' for English and Mathematics for 'successful' modern foreign language and physical science candidates.

Subject	'Group Ratings'		
	Physical science	Modern foreign languages	*SQA National Rating*
English	-0.27	0.48	0.26
Mathematics	-0.13	-0.92	-0.76

These figures again suggest that 'subject difficulty' is relative and personal, a subject does not have the same 'difficulty' for everyone, so why is a **single** *National Rating* used?

CONCLUSIONS

The implications of this investigation are quite fatal for the *Relative Rating* indicator. A Mathematics department in a school in an affluent neighbourhood, with a larger than normal number of senior boys hoping to become engineers or scientists, could expect its *Relative Rating* to be 'notably' positive. In a neighbouring school, with an unusually large number of girls applying for university courses in English or Drama, the Mathematics department might obtain a 'notably' negative rating. The difference between the two departments could be as much as one whole grade, yet in neither case has the rating anything to do with the 'effectiveness' of the departments themselves, it is simply a result of the erroneous assumptions made when the National Rating is included. Unless schools are able to undertake this kind of detailed analysis of their student profiles, in order to correct for such anomalies, they would be advised to ignore their *Relative Ratings* entirely.

Fortunately, most students do ignore them. There is absolutely no significant correlation between the *Relative Rating* that a department receives and the popularity of the subject it offers (as measured by the proportion of the school's candidates that undertake that subject). We might surely expect 'successful' departments with 'above average' performance in public examinations to attract more candidates. The fact that they don't, indicates that these performance indicators aren't measuring anything useful in the first place.

The way that the SEED uses the *Relative Rating* indicator assumes that the random factors that contribute to candidates' examination performance cancel one another out when the results are averaged. The evidence from our investigations is that this is only partially true and that a significant proportion of the difference between a subject department's average results and the national average is still due to these random factors. Only a small proportion of the differences between departments in their candidates' examination achievements can be identified as the influence of the school and not all of this is actually under the control of the school or their departments – it may well be due to external 'social context' factors.

This is not to say that schools have little effect on candidates' examination results, they obviously do, as a comparison between pupils, who do or do not attend school, clearly demonstrates. But the *Value Added* and *Relative Rating* indicators merely compare each department with the average, and the resulting figures only demonstrate that schools, generally, have almost the *same* effect on all their candidates, once the social differences between the latter have been allowed for.

The performance of candidates is inconsistent, and some of the *School Rating* is the remaining inconsistency when results are averaged, which makes it uncertain. The calculation of a subject department's *Relative Rating* from its *School Rating* depends upon the subject's *National Rating* being a reliable figure. Our study demonstrates that it is just the average of yet another random process, such that the relative 'difficulty' of the subject depends upon the candidate's 'ability', sex and personal interests. The use of a single *National Rating* to 'correct' each subject department's *School Rating* can only work if that department's candidates are exactly representative of all that subject's candidates. If there were a thousand candidates in each department, who were randomly selected, this might be true, but most schools have only a few dozen candidates in each department who are largely self selected. In these circumstances, the *Relative Rating* is not just uncertain, it is completely meaningless.

None of this should be taken to imply that there are no differences between schools or between departments or even between teachers. Personal experience (exemplified in the testimonies given to their teachers by prominent people in the Tuesday Guardian) asserts that good teachers do make a difference. If the candidates in a subject department in one school attain a mean range-score of 6.0 in that subject and they gain a mean range-score of 7.0 in their other subjects, this is not necessarily because this department has done a good job in preparing its candidates for the examination. It may be that these candidates find this subject 'easier' (whatever that might mean!) or it may be because they have a greater motivation to study this subject. On the other hand, it may be because they have excellent teachers, and that is the problem, we just don't know. And the department's *Relative Rating* certainly won't tell us!

REFERENCES

Bardell, G.S., Forrest, G.M. and Shoesmith, D.J. (1978) *Comparability in GCE: a Review of the Board's Studies 1964-1977*. Manchester: JMB.

Coe, R. and Fitz-Gibbon, C.T. (1998) School Effectiveness Research: criticisms and recommendations. *Oxford Review of Education,* **24**, 4, pp. 421-438.

Forrest, G.M. and Vickerman, C. (1982) *Standards in GCE: subject pairs comparisons, 1972-80* . Manchester: Joint Matriculation Board.

Fitz-Gibbon, C.T. (1992) *School Effects at A-level: Genesis of an Information System?* Newcastle upon Tyne: CEM.

Fitz-Gibbon, C.T. and Vincent, L. (1994) *Candidates' Performance in Public Examinations in Mathematics and Science*., SCAA Report, Newcastle-upon-Tyne, CEM.

Fitz-Gibbon, C.T. (1996) *Monitoring Education: Indicators, Quality and Effectiveness,* London: Cassell.

Fitz-Gibbon, C.T. and Vincent, L. (1997) Difficulties regarding subject difficulties: developing reasonable explanations for observable data. *Oxford Review of Education* **23**, 3, pp. 291-298.

Forrest, G.M. and Vickerman, C. (1982) *Standards in GCE: subject pairs comparisons, 1972-80* , Manchester: Joint Matriculation Board.

Gibson, A. and Asthana, S. (1998) Schools, Pupils and Examination Results: contextualising school 'performance', *British Educational Research Journal*, **24**, 3, pp.269-282.

Goldstein, H. and Cresswell, M.J.(1996) The Comparability of Different Subjects in Public Examinations: a theoretical and practical critique. *Oxford Review of Education* **22**, 4, pp. 435-442.

Goldstein, H. and Spiegelhalter, D.J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance *J.R.Statist. Soc*. **159**, 3, pp.385-443.

Gray, J., Allnutt, D., Gardner, J., Blackham, C. and Frost, B. (1995) *The development of a National Framework for estimating value-added at GCSE A/AS Level - Briefing for Schools and Colleges,* London: DfEE.

Gray, J. (1996) *The use of assessment to compare institutions.* in Goldstein, H. and Lewis, T. (Eds.) *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley & Sons.

Kelly, A. (1976) 'A study of the comparability of external examinations in different subjects *Research in Education,* **16**, pp.50-53.

Kendall, L. (1995) Contextualisation of school examination results, 1992, *Educational Research,* **37**, pp.123-139.

Mortimore J and Mortimore, P. (1984) *Secondary School Examinations* London: Heinemann.

Murphy, R. (1997) Drawing outrageous conclusions from national assessment results: where will it all end? *British Journal of Curriculum and Assessment*, **7**, pp.32-34.

Nuttall, D.L., Backhouse, J.K. and Willmott, A.S. (1974) *Comparability of Standards between Subjects*., Schools Council Examinations, Bulletin 29, London: Evans/Methuen Educational.

Nuttall, D.L., Goldstein, H., Prosser, R. and Rasbash, J. (1989) Differential school effectiveness. *International Journal of Educational. Research*, **13**, 7, pp.769-776.

O'Donaghue, C., Thomas, S., Goldstein, H. and Knight, T. (1997) *1996 Study on Value Added for 16-18 year olds in England*, London: HMSO.

Smith, D. J. and Tomlinson, S. (1989) *The School Effect: A Study of Multi-Racial Comprehensives*. London: Policy Studies Institute.

SOEID (1991) *Using examination results in school self-evaluation: Relative Ratings and National Comparison factors*. Edinburgh: HMSO.

SOEID (1998) *Taking a closer look at attainment in secondary schools* in *Raising Standards – Setting Targets,* Edinburgh, HMSO

Sparkes, R.A. (1999) Value Added – an uncertain measure, *Scottish Educational Review*, **31**, 1, pp. 21-34.

Thomas, S and Mortimore, P. (1996) Comparison of VA models for secondary school effectiveness*, Research Papers in Education*, **11**, 1, pp.5-33.

Thomas, S., Smees, R.T., MacBeath, J., Sammons, P., Robertson, P. and Mortimore, P. (1998) *Creating a Value-Added Framework for Scottish Schools*. Policy Paper No 2, QIE, Glasgow: University of Strathclyde.

Tymms, P.B. and Fitz-Gibbon, C.T. (1991) 'A Comparison of Examination Boards: A levels'*, Oxford Review of Education*, **17**, 1, pp.17-32.

Tymms, P.B. and Fitz-Gibbon, C.T. (1990) *The stability of school effectiveness indicators,* Publication 32, Newcastle: CEM, University of Newcastle upon Tyne.

Woodhouse, G. and Goldstein, H. (1996) *The Statistical Analysis of Institution-based Data* in Goldstein, H. and Lewis, T. (Eds.) *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley & Sons.