VALUE ADDED - AN UNCERTAIN MEASURE

R.A.SPARKES

SYNOPSIS

School performance indicators based on 'raw results' are being superseded by those based on 'value added' methods. In Scotland, one Value Added indicator used in secondary schools measures students' relative 'progress' between Standard and Higher grade. The average 'progress' of a school's candidates in a particular subject is that department's Value Added indicator, but this fluctuates from year to year. The random factors that affect students' performance (over which schools have little or no control) were simulated on a computer to determine the size of this fluctuation. It was found that the confidence intervals for most subject departments are so wide that it becomes almost impossible to distinguish between them. The conclusion is that this indicator is an uncertain way of measuring 'effectiveness'.

WHY 'VALUE ADDED'?

The grade obtained in a public examination by a particular candidate depends upon many contributory factors. At one extreme, some factors are completely random and cannot be influenced by candidates or their schools; for example, the uncertainty in grading the candidate's answers in the examination (Nuttall and Willmott, 1972; Mortimore and Mortimore, 1984). At the other extreme, the factors are wholly dependent upon the school; such as whether the teacher actually taught a topic in the syllabus. In between, factors depend partly on candidates and partly on their teachers, parents, peers and other social circumstances. For purposes of analysis, it is convenient to divide these factors into two categories, according to whether they appear to be common to the candidates of a particular subject department in a school (hereinafter called the 'School Effect') or whether they do not (Random Factors). The 'School Effect' is measured by the *between school* variance in candidates' performance, whilst the Random Factors are determined by the within school variance.

Annually, 'league tables' of the public examination performance of candidates in different schools are published in the national and local press. These are strongly criticised for being 'unfair', because some of the factors that contribute to candidates' performance are not under the control of the school (for example, the socio-economic status of a candidate's parents). There has been much criticism of some performance indicators because they do not take these 'contextual factors' into account (Smith and Tomlinson, 1989; Kendall, 1995; Murphy, 1997), which is why *value added* methods are being introduced (Woodhouse and Goldstein, 1996).

In the value added method, allowance is made for contextual factors by using candidates' prior performance in a public examination as a covariate, which then concentrates attention on the '*progress*' made by each candidate thereafter. Because all candidates make some progress, it would be more accurate to call it their *relative progress*; they are just being compared with the average of *all* candidates. The average 'progress' made by the school's candidates in one particular subject is called the *Value Added* (VA) indicator of that department and this is "widely regarded as providing more accurate measures of school effectiveness than the raw results" (Thomas, *et al*, 1998). Some researchers (Fitz-Gibbon, 1995a; Jesson, 1996; Thomas & Mortimore, 1996) regard socio-economic factors as of minor importance after 'prior performance' has been taken into account, but others (Gibson and Asthana, 1998a, 1998b) dispute how far it does compensate for the social context of the school and there does now seem to be some attempt to include this in the VA indicator as well (Thomas *et al.*, 1998).

The debate about contextual factors centres on the *validity* of value added indicators for measuring school effectiveness. While not denying the extreme importance of this, the current study is more concerned with the reliability of the VA indicator. The apparent assumption seems to be that, when the examination results for a particular subject in a school are aggregated to form its VA indicator, the Random Factors cancel one another out and only the 'School Effect' remains. The researcher's background in physics led him to question this assumption and to investigate the *accuracy* of the VA figures that schools were being given.

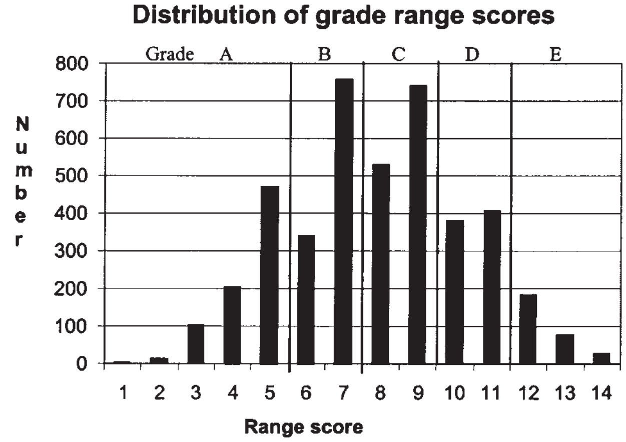
HOW THE VA INDICATOR IS CALCULATED IN SCOTLAND

In calculating the VA indicator, 'prior performance' is measured by the mean grade obtained by each candidate in his or her Standard grade examinations the previous year - called the Standard grade GPA (grade point average). Research in the rest of the UK consistently shows that the GPA at GCSE/O-level is the best predictor of a candidate's GCE A-level result in any particular subject, better even than performance in that same subject at GCSE/O-level (Kelly, 1976; Fitz-Gibbon and Tymms, 1991; Fitz-Gibbon, 1992, 1995b, 1996; Gray, *et al*, 1995; O'Donaghue, 1997).

To illustrate the methodology of the SOEID, the VA indicator for one particular subject, Geography, is calculated below. Geography was chosen for this investigation because it is both an Arts and a Science based Higher subject; it is about average in terms of the number of its candidates, it has the same *difficulty* as most other Higher subjects and its correlation with GPA is similar too. In Scotland, the Standard grade examination awards are made by the Scottish Qualifications Agency (SQA) in 7 grades. A year or more later, candidates are presented for the Higher grade examination, which is awarded at grades A, B, C, D and E. This examination is actually assessed on the numerical scale 1 to 14 (called range scores) which are then converted to letter grades according to the code: scores 1, 2, 3, 4 and 5 are Grade A; scores 6 and 7 are Grade B; scores 8 and 9 are Grade C; scores 10 and 11 are Grade D and scores 12, 13 and 14 are Grade E. It should be noted that, for both examinations, a lower score represents a *higher* attainment.

It seems unlikely that a simple conversion system is used to turn examination marks into range scores, but rather that the latter are chosen with some reference the letter grades. This can seen from the distribution of range scores amongst the Higher Geography candidates (Figure 1).

Figure 1

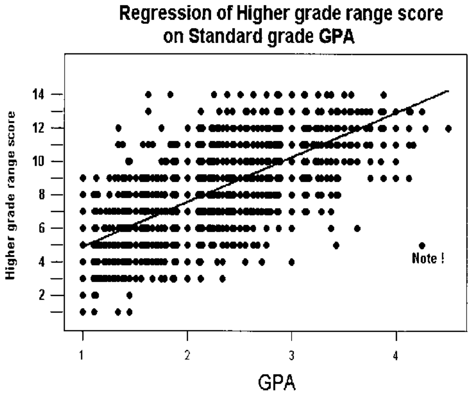


It appears that some candidates at the top of a particular band are ‘promoted’ to the next band, so that the number of 5s (grade A) is higher than expected from a ‘normal distribution’ and the number of 6s (grade B) is lower. The same is true for the interface between grades C and B and between grades D and C. Despite this non-normality, because range scores represent a finer division, they are used in place of the letter grades when calculating VA indicators. The results of this calculation are then halved to give figures that can be interpreted in letter grades, rather than range scores. The justification for this conversion is dubious, although it makes no difference when VA indicators are used to determine whether departments are above or below average.

For this investigation, a full set of results was obtained from the SQA and a sample of all candidates, who had been presented for Higher Geography in 1997, was extracted from it. It was necessary to remove from the sample all schools with fewer than 10 candidates, because it has been found that the VA indicators of subject departments with fewer than 10 candidates are too inconsistent and the SOEID does not produce VA indicators for such schools (SOEID, 1993). When this was done, 199 schools (about half of the secondary schools in Scotland) remained, (4232 candidates).

To determine the regression line (Figure 2), the actual results that *all* 1997 Higher grade Geography candidates had achieved were plotted against the same candidates’ 1996 GPAs.

Figure 2



The best straight line through these points (ordinary least squares regression) was drawn and used to form the regression equation, which, for these results is given by Eq. 1.

Predicted range score in Higher Geography of candidate$_i$ = 2.20 + 2.76 x GPA$_i$ ______Eq. 1

Regression analysis shows that the GPA accounts for 52.4% of the variance in candidates’ results (from the adjusted square of the correlation coefficient - R^2). Thus, although ‘prior performance’ is a very useful predictor of Higher grade success (probably the best single predictor there is), other factors, including the ‘School Effect’, still have some influence – contributing 47.6% to the variance.

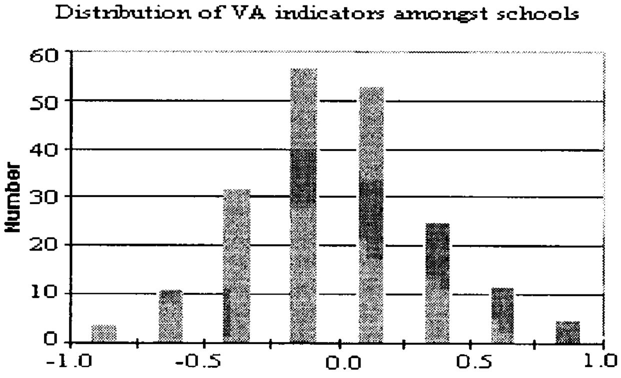
This regression model is used to determine the grade that candidates with a particular GPA might be expected to achieve in Higher Geography. The difference between this expectation and the candidate’s actual Higher grade result is called a *residual*; however, because higher attainment has a lower range score, the residual is defined in reverse, as follows:

residual = predicted range score - actual range score ______Eq. 2

This ensures that candidates with positive residuals have done better than other candidates with an equivalent GPA and candidates with negative residuals have done worse.

To obtain the VA indicator of a school's Geography department, the *mean residual* of all the Geography candidates in that school is calculated and then halved, so as to produce an indicator in terms of letter grades. For Higher grade Geography in 1997, the VA indicators of different departments (Figure 3) were found to range between -1.0 and +1.0, the value of +0.5, for example, indicating that this department's candidates in Geography achieved *half a grade* better than the national average for candidates of 'similar ability'.

Figure 3



This chart shows that candidates of 'similar ability' attained an A grade, a B grade or a C grade depending on which school they attended. Stated in these terms, the case for publishing 'performance tables' is understandable.

THE RELATIONSHIP BETWEEN THE VA INDICATOR AND DEPARTMENTAL 'EFFECTIVENESS'.

The assumption in this value added model is that the mean residual is due to departmental 'effectiveness'. When the mean residual is calculated, it is assumed that any *random* factors which might have influenced candidates' results cancel one another out, so that only those factors that are common to all the department's candidates remain (i.e. the 'School Effect'). On this basis, the mean residual is a direct measure of departmental effectiveness, untainted by the contextual factors that influence 'raw results'. This explains why the VA indicator is so appealing, it appears to give a *scientific* way of categorising schools (Gibson & Asthana, 1998b).

It is true that, in the nation as a whole, the candidates of a particular subject are likely to have different levels of certain influential factors (for example, intelligence and motivation) which are randomly distributed and which cancel one another out when averaged. However, this is less likely to be true within a school subject department, because of the much smaller number of candidates. One particular Higher Geography class might have a larger than usual number of 'intelligent' pupils in one particular year, so that it's VA indicator is enhanced even if the 'effectiveness' of its teachers is identical to previous years. It follows that a department's VA indicator will fluctuate from one year to the next, depending on whether it had 'good' or 'bad' students (to use teachers' own terminology).

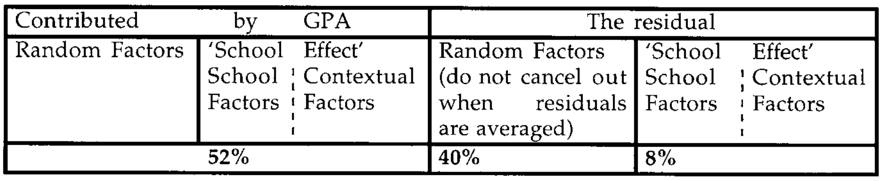
This unpredictability can be seen in the Geography results shown in Figure 2, in which attention is drawn to the student with a range score of 5 (grade A) in the Higher Geography examination, but who is predicted to obtain a range score of 14 (grade E) because his Standard grade GPA was only 4.25. His residual is thus +4.5 grades (+9 range scores), so if he had been in a small department, his residual would have had a very large influence on that department's 1997 VA indicator, raising it by nearly half a grade. Although most candidates' residuals are smaller than 4.5, they are still sufficient to cause the VA indicator to fluctuate widely from year to year.

If the VA indicator *does* measure the 'value added' to a candidate's Higher grade examination result after taking into account his or her prior performance, then it should be possible to improve the regression model by including this indicator as an additional input variable (SchoolVA), (following 'model 5' developed by Aitkin and Longford, 1986) whereupon we should be able to predict the Higher grade results with more certainty. When this is done for our Geography results, the following regression equation is produced.

$$\text{Predicted range score} = 2.25 + 2.65 \text{ x GPA} - 1.00 \text{ x SchoolVA} \qquad \_\_\_\text{Eq. 3}$$

In this equation, all three terms were significant at the 0.05 level (the negative coefficient for the SchoolVA is because of the way the residual is defined in Eq. 2). The value of R^2 for this equation was 60.4%, showing it to be a more accurate model than Eq. 1. Analysis of variance showed the GPA term to explain 52.4% of the variance (exactly as before) and the SchoolVA term an additional 8.0%. In other words, the SchoolVA term contributes only 17% to the variance in the candidates' residuals. Similar figures have been obtained by others for A-levels (Willms, 1987; Reynolds, 1991; Cuttance, 1991; O'Donoghue *et al,* 1997).

Figure 4 The variance in candidates' Higher grade results



Contributed by GPA			The residual		
Random Factors	'School Effect' School Factors	Contextual Factors	Random Factors (do not cancel out when residuals are averaged)	'School Effect' School Factors	Contextual Factors
52%			40%	8%	

This implies that the variance due to our Random Factors (83%) far exceeds that due to the 'School Effect' (17%), which either means that the VA indicator is a poor measure of the 'School Effect' or that schools are not particularly effective. The latter is a doubtful interpretation, though, because candidates presented for the Higher examinations have already been in their schools for the four years prior to S5. It would be ridiculous to assert that these schools had had no influence during these years so there is almost certainly some 'School Effect' in the candidates' GPAs, which does not show up in the residual. Furthermore, the VA indicator only measures *relative* effectiveness; if all departments have the same effect on their candidates, they make no contribution to the variance between candidates.

It is worth noting that even this 17% contribution of the 'School Effect' to the variance in the residuals cannot be wholly attributed to a department, because it includes contextual factors as well as school factors. This can also be seen from the correlation between 'raw' results (based on the average range score achieved in each department by its candidates) and the VA indicators of these departments, which yields the relatively high value of -0.63 (negative, because of the way the residual is defined).

These observations lead to the model shown in Figure 4 to account for the variance in candidates' Higher grade results.

It is this enormous contribution of the Random Factors to the candidates' residuals that renders the VA indicator almost worthless as a measure of departmental effectiveness.

THE SOEID'S INTERPRETATION OF THE VA INDICATOR

The SOEID has been making VA indicators available to schools for several years and has published guidelines for school boards and head teachers on how they might properly be interpreted (SOEID, 1993, 1997a, 1997b). The underlying assumption is that, once 'prior performance' has been entered into the regression equation, any resulting difference between candidates' residuals must be due to the 'effect' of their subject departments. The guidelines do not specifically describe the VA indicator as showing the *effectiveness* of the subject department in the school concerned, but rather that they show whether its *candidates* are "better or worse than the national average of candidates with their particular GPA" (SOEID, 1993). Even so, there is an implication that it does measure something that is due to the subject department itself.

> Positive value added indicators suggest that the actual attainment of the pupils presented for Higher grade was better than would have been predicted for this group of pupils given what was known about their prior attainment at Standard grade and the relationship between Standard grade and Higher grade performance nationally in the year concerned. We can conclude that this department is *doing a better than average job* of preparing its pupils for examination. (emphasis added) (SOEID, 1993, p.8/13).

This interpretation is invalid because there is no indication of the *measurement error* in the VA indicator. In one case, two indicators are given as -0.24 and -0.25 and the guidelines comment:

> The value added indicators for English suggest that performance was weak in 1992 and notably weak in 1991. (The difference in performance over the two years, 0.01, was extremely marginal, however.) (SOEID, 1993, p.8/16)

Only if the measurement error is about 0.01 could such a difference be described as 'marginal' and, if the error is greater than 0.01, distinguishing between these two indicators is invalid. Of course the *number* 0.25 is bigger than 0.24, but these are *measurements* and they have associated inaccuracy. If the figures are to be used to infer a school's effectiveness, then the size of this inaccuracy needs to be stated before they can properly be interpreted.

In the same guidelines, an example is provided (SOEID, 1993, Table 8.1) of a subject department that achieved a VA indicator of 0.15. In the commentary it is stated that "this department is doing a better than average job of preparing its pupils for examination" (p.8/13). Another example shows a subject department that achieved a VA indicator of -0.10. In the commentary it is stated that "this department is doing a worse than average job of preparing its pupils for examination" (p.8/13). In some cases, VA indicators (for example, those above 0.25 or below -0.25) are asterisked and described as "notably better or worse than average" and the commentary on the VA indicators for English, quoted above, distinguishes between 'weak' (-0.24) and 'notably weak' (-0.25), so there is some hint here of a criterion level of 0.25 for 'notable'. However this is not given as a measurement error and it is nowhere stated, for example, that absolute values below 0.25 are insignificant and may be ignored.

How Good are our Results (SOEID, 1997b) conveys the same message, where an example is shown of a department's VA measures over three years. Again, absolute values exceeding roughly 0.25 are marked as 'notable'. Interestingly, though, the figures given for these schools for the other two years varied by at least this amount. Surely, if a subject department's results really are 'notable', then they ought to be above the level of random fluctuation. The statement that added value "provides a useful method for comparing departments against the national average" (p.2) is invalid—the comparison of means cannot be made meaningfully without stating their associated errors.

In the SOEID guidelines on other performance indicators (SOEID, 1991), there is an explanation of how the figures for *Relative Ratings* (a performance indicator which compares the results of candidates of different subjects in the same school) are to be interpreted, which sheds some light on their interpretation of VA indicators.

> The extent to which a relative rating differs from zero is a measure of a department's relative strength or weakness.
>
> **Threshold values** are identified. Relative ratings beyond these values suggest variations in performance which are notable. They are highlighted with an asterisk for convenience.
>
> The procedure for highlighting relative ratings with an asterisk has an empirical and a statistical basis. If a relative rating is to be highlighted we must be sure beyond any reasonable doubt that its deviation from zero is not a random fluctuation but represents an actual deviation from the mean. (p. 7/18).

The same guidelines show how their threshold level for a 'notable' relative rating is determined by exactly the same method used by statisticians to determine a 95% confidence level. The standard deviation of candidates' relative ratings is calculated and divided by the square root of the number of candidates in the department to produce a standard error in the department's relative rating. This is then multiplied by 1.96, to give the threshold value (SOEID, 1991, p. 7/29).

There is no dispute with this calculation, but only with its subsequent interpretation. In terms of the VA indicator, it seems that any value is worth reporting since it indicates whether a department is above or below average. If the figure exceeds its threshold value, then it is 'notably' above or below average. This is not, though, the usual way of interpreting such figures. The statistical way of determining whether the mean of a set of data is significantly different from zero is to calculate how much greater it is than the standard error in this mean. Suppose that a department's VA indicator is measured as -0.15 and its standard error is 0.1. Assuming that VA indicators are normally distributed (and Figure 3 suggests that they are), then this department's 'true' VA indicator has a 19 out of 20 chance of being in the range -0.35 and +0.05 (the 95% confidence interval for this measurement). That is, the 'true' VA indicator could have been zero (which would indicate that this department is about average). The 95% confidence interval is the mean plus or minus two (actually 1.96) standard errors and it would be normal practice in statistics to say that any confidence interval that included zero is *not significantly different* from zero. The guidelines ought to state that departments with VA indicators which are **not** asterisked are **indistinguishable from the average**.

In the more recent *Raising standards - setting targets* (SOEID, 1997c), there is a possible change of emphasis. A VA indicator of -0.1 is now described as "broadly in line with the national expectation" (p.12) and two subject departments identified as "better" or "less well" had figures of 0.75 and -0.7 of a grade respectively. These figures are larger than in previous examples, and indicate, perhaps, a reversion to the more usual interpretation of statistical significance.

The SOEID is not ignorant of the effect of random fluctuation on the VA indicator (as the relative ratings quotation above reveals). The figures sent to schools show *three successive years* at a time, to give some indication of the annual variation in the figures (SEB, 1996; SOEID, 1997a). There are also repeated warnings that the figures should only be interpreted in the light of the school's own particular situation. Our criticism is of their failure to specify a confidence interval. Similar criticisms have been made of the DfEE (Goldstein, 1998a).

> …it is now accepted practice, when publishing value added scores (or bands) to include 'confidence intervals' so that instead of a single number a range of plausible values is given. The government is fully aware of this issue yet chooses not to even mention it. In so doing it is withholding crucial information from readers to the point of actually misleading people.

STANDARD ERROR IN THE VA INDICATOR

The normal way of calculating its confidence interval would be to measure a VA indicator over a number of years and to determine its standard deviation, although departments and examination courses are rarely consistent enough for this to be satisfactory. Goldstein and Spiegelhalter (1996) and Goldstein (1998b) did this for various performance indicators, both 'value added' and 'raw results', and found that their confidence intervals were larger than the differences between most schools, which meant that the published figures were of little value for their intended purpose. Other research has also shown that VA measures are not consistent within a single department from year to year, particularly when the number of candidates is small (Tymms and Fitz-Gibbon, 1990; Fitz-Gibbon, 1996, Nuttall *et al,* 1989); they vary markedly from one subject to another and are quite different for different groups of pupils in the same school (Fitz-Gibbon, 1996). We showed above that most (83%) of the factors that influence a candidate's residual are unexplained and random.

An attempt was made to determine the size of the annual fluctuation in the VA indicator in just this traditional way. A sample of about 250 schools was obtained from the SOEID with their value added indicators listed for the three years 1996, 1997 and 1998 for each of the subjects for which candidates were presented. The Geography candidates were selected from this list and departments with fewer than 10 presentations were eliminated, which produced three value added indicators for each of 147 Geography departments (441 indicators overall). In confirmation of the model shown in Figure 4 and the research described above, the VA indicators were not consistent. The standard deviation was determined for each of the departments and their (root mean square) average value was 0.20. Six departments had all three indicators significantly above average and five departments were consistently, and significantly, below average, which suggests that 'School Effects' were probably also influential in a few instances. However, since the number of candidates in these departments was not known, it was not possible to pursue this investigation further. In any case it is not very informative with only three sets of data for each department.

Standard grades only became the norm by 1994 and Higher grades may not be the same after 1999, so the prospect of ever obtaining sufficient data to carry out this analysis is unlikely. In any case, departments do change from year to year (principal teachers retire or are appointed all the time) and it would be impossible to distinguish between fluctuations caused by changes in the 'School Effect' and those produced by random changes in the candidates. What is needed is a set of departments which remain unchanged over many years.

Tymms (1996) has investigated models, which give a mathematical description of the chaotic processes that appear to be involved in candidates' examination results.

He noted the unpredictability of A-level results and indicated what effect this might have on a school's performance indicators:

> the simulation models would suggest that even if it were possible to arrange for exactly the same class to have exactly the same teacher for two years in the same classroom living through the same two years that the outcomes would not be the same. (p. 133)

This pointed to one possible approach – the simulation of 25 years of VA indicators for each of the 199 schools in our sample, using the model constructed in Figure 4, where the candidates' GPAs and the 'School Effect' were identical from year to year.

To do this, it was decided to create a set of simulated Higher Geography results that was as accurate as possible. Some researchers use a curvilinear regression model to make their calculations of the VA indicator more precise (O'Donaghue, *et al*, 1997), although Fitz-Gibbon (1995b, 1996) maintains that, for determining a subject department's VA indicator, the linear model is nearly as accurate as the curvilinear model, and is very much simpler for non-statisticians to understand.

Since we were aiming for the greatest accuracy in our simulated results, it was decided to use the curvilinear model to create them. Many equations were tried, but the best was found to be as follows:

$$\text{Predicted range score} = -0.03 + 4.88 \text{ x GPA} - 0.50 \text{ x GPA}^2 - 1.01 \text{ x SchoolVA} \quad \_\_\text{Eq. 4}$$

This regression equation accounted for 61.5% of the variance, an improvement of only 1.1% overall, giving some support to Fitz-Gibbon's view that the curvilinear model has no great advantages for calculating the VA indicator.

The simulated data set based on this equation assigned to each Higher Geography candidate the same Standard grade GPA that they had obtained in reality and assigned to their departments the same SchoolVA that they had actually obtained in 1997, so that the range score in Higher Geography of candidate$_i$ was given by

$$\text{Range score} = -0.03 + 4.88 \text{ x GPA}_i - 0.50 \text{ x GPA}_i^2 - 1.01 \text{ x SchoolVA} + \varepsilon_i \quad \_\_\text{Eq. 5}$$

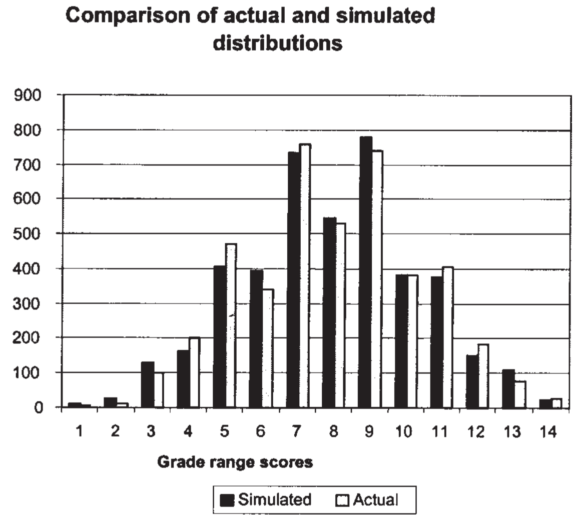
which is the curvilinear regression line determined above (Eq. 4) with the addition of ε_i to represent the Random Factors. ε_i was a normally distributed random number with a mean of 0.00 and standard deviation of 1.65, produced by adding up a set of 32 random numbers, each between -1.000 and +1.000. It was devised to contribute the remaining 40% to the variance in the range scores (once the variance due to the GPA and the 'School Effect' had been taken into account), and to ensure that Eq. 5 reproduced the original actual Higher Geography scores as closely as possible (i.e. with the same mean, standard deviation and distribution, even taking into account the predominance of odd numbered scores over even-numbered as shown in Figure 1). Each range score was rounded to its nearest whole number, with the numbers 6, 8 and 10 being rounded down to 5, 7 and 9 slightly more generously, so as to replicate the distribution of the actual examination range scores. This equation was applied to each of the 4232 candidates in each of the 199 schools of the original sample.

The success of the simulation model is shown in Table 1, which compares the means and standard deviations (and their Pearson correlation coefficients with candidates' GPAs) for the simulated data and for the actual data, and also in Figure 5, which compares their distributions.

Table 1. Comparison between the actual and the simulated Geography range scores

	Actual H-grade	Simulated H-grade
Mean	7.93	7.93
Standard deviation	2.40	2.43
Correlation coefficient with candidates' GPA	0.72	0.72

Figure 5



Linear regression analysis was applied to the simulated results in the same way as with the actual results. The regression equation was found to be

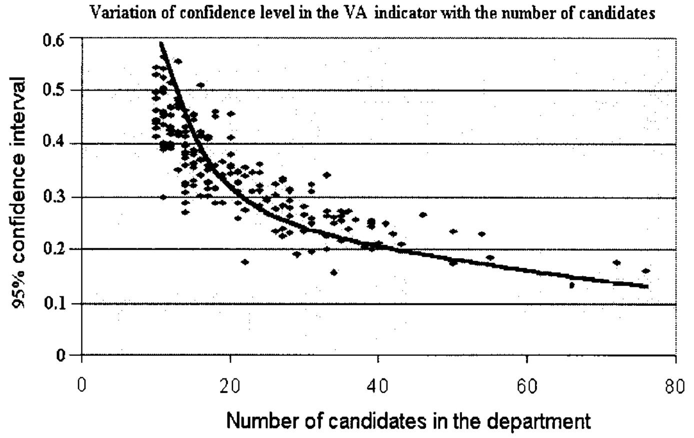
Predicted range score in Higher Geography = 2.23 + 2.66 x GPA - 1.00 x SchoolVA

which may be compared with Eq. 3. The candidate's GPA was found to contribute 50.7% of the variance in the candidates' simulated range scores (actual results 52.4%), the school's VA indicator contributed 10.0% (actual results 8.0%) with the overall model explaining 60.7% of the variance (actual results 60.4%). The simulated VA indicators of the departments were very highly correlated (0.99) with the actual VA indicators. The simulated results thus showed sufficient resemblance to the actual results that the generating equation (Eq. 5) could be confidently used to simulate 25 successive years with the 'same' pupils and teachers.

These 25 sets of simulated results revealed a fluctuation in each department's VA indicator (the values being reported in grades). The VA indicators of the 'least effective' department varied from -0.3 to -1.3, a range of 1.0, and the 'most effective' school varied from +0.6 to +1.2, a range of 0.6. The standard deviations of each school's set of VA indicators varied from 0.07 to 0.28, with a root mean square average of 0.19, giving a 95% confidence level of 0.38 grades. Considering that the standard deviation of the actual VA indicators across the Geography departments in the sample (Figure 3) was 0.35, this suggests that a school's position in any 'Value Added league table' is likely to fluctuate markedly from year to year, solely because of random factors over which it has no control.

When the 95% confidence level (1.96 standard deviations) in a subject department's simulated VA indicator is plotted against the number of its candidates, the influence of randomness on small departments is particularly apparent (Figure 6).

Figure 6



The 95% confidence interval for the VA indicator of subject departments with 21 candidates (the national average) was 0.34 but for departments with just 10 candidates rose to nearly 0.6. This phenomenon confirms the investigations of Tymms and Fitz-Gibbon (1990) into the stability of the VA indicator for Advanced level Mathematics, where they noted that its reliability was very dependent upon the number of candidates in the department. The decision of the SOEID not to produce VA indicators for departments with fewer than ten candidates is thus fully justified.

With this 95% confidence interval, two average-sized departments with VA indicators closer than 0.38 cannot be considered 'different' and, from our figures, it is therefore impossible to distinguish between more than half the schools in the country. This is not a new discovery; it confirms the observations made by most other researchers in this field. Goldstein (1998b) states:

> Up to a point these models also allow us to identify the contribution from individual schools—their so called 'value added scores'. Yet we now know from a number of studies that estimates of these contributions have so much statistical uncertainty attaching to them that it is impossible reliably to make valid comparisons between most schools… schools can only be separated statistically if their intervals do not overlap… (p. 7)

Thomas and Mortimore (1996), in an investigation of value added measures, used a 95% confidence interval with over 100 schools but found only a few that differed from each other by more than the expected amount. They commented

> In comparing the confidence intervals for any two schools… it can be seen that the majority overlap. This indicates that one cannot safely, or confidently, differentiate between the performance of these schools. As a 'conservative' guide, only those schools with non-overlapping ranges have clearly different results (p.14)

Finally, Gray (1996) comments:

> The main substantive conclusion to be drawn from the analyses which have been conducted to date is that the considerable majority of schools achieve precisely the sort of results one would predict from knowledge of their intakes. A few may do substantially better while a similarly small number may do substantially worse… The recent use of procedures for estimating the "uncertainty" attached to the estimates of "effectiveness" for specific schools has reinforced the view that a school's effectiveness is not a precisely estimable quantity… Where such "bands" (of 'uncertainty') have been constructed they tend to show that the performances of some two-thirds to three-quarters of schools cannot properly be distinguished from one another. (p. 132-3).

THE CONFIDENCE INTERVAL

The purpose of this investigation was to determine the uncertainty in a subject department's VA indicator, so that when departments look at the figures supplied to them by the SOEID, they might better be able to judge whether they need to be concerned with their quality. Our study of three sets of real VA indicators suggests a 95% confidence level of around plus or minus 0.4 grades. The simulation suggests plus or minus 0.34 grades for an average department, but ranging from 0.6 for a department with ten candidates to 0.15 for a department with more than 60 candidates.

There is, though, an unresolved argument about what percentage level of confidence should be adopted. The SOEID uses a 95% confidence interval, which means that there is a 1 in 20 chance that a department's VA indicator will be marked as 'notable', even when that department is exactly in line with the national average. With over 400 schools and at least 5000 departments, this means that about 100 departments could be unfairly designated as 'notably' below average, when they are not. If this designation were to be published in the national press, the damage done to the departments' reputations and their teachers' self esteem could be enormous (yet completely undeserved).

For regression analysis to work properly, the dependent variable (in this case, the candidate's Higher grade range score) should be based on an interval scale (with the same 'distance' between range scores 1 and 2 as between 6 and 7, which is unlikely). Grades and range scores are not continuous either, with the result that candidates with GPAs less than 4.5 must necessarily achieve positive residuals if they sit any Higher examinations at all; they can never do worse than fail! There is a 'ceiling' to the GPA which also produces 'unusual' residuals (any candidate with a range score of 1, 2, 3 or 4 *must* gain a positive residual, whatever their GPA). Finally, there is the use of a linear regression equation, when a curvilinear one gives more accurate results. None of these restrictions invalidates the method of calculating the VA indicator, but they do suggest that a 95% confidence level may be rather too low.

We could reduce the chance of unfairly designating a satisfactory department as 'below average' by setting the confidence level higher, at say 99% or 99.7% (three standard deviations), but this increases the possibility that an unsatisfactory department will be designated as 'average'. There does not seem to be any way of resolving this problem, which itself suggests that the VA indicator is a poor way to determine a department's effectiveness.

CONCLUSION

An average sized subject department with a VA indicator below -0.34 will not know whether this is because they are "doing a worse than average job of preparing their pupils for examination" or because they happen to have had a "bad" set of candidates. Smaller departments require even larger negative values before becoming concerned with their performance. Furthermore, a department's VA indicator still contains a large proportion of 'contextual factors', so the department could only be identified as 'ineffective' if it was known how the school's social context had affected its value.

Precisely what does the VA indicator indicate?

REFERENCES

Aitkin, M. and Longford, N. (1986) Statistical Modelling Issues in School Effectiveness Studies *J.R.Statist.* Soc. 140, (1), 1-43.

Cuttance, P. (1991) Assessing the effectiveness of schools in Reynolds, D. and Cuttance, P. (eds.) *School effectiveness*, London: Cassell.

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

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

Fitz-Gibbon, C. T. (1995a) *The Value Added National Project: issues to be considered in the design of a national value added system*, London: SCAA.

Fitz-Gibbon, C. T. (1995b) 'A level results in Comprehensive Schools: The COMBSE Project', *Oxford Review of Education*, 11 (1), Oxford: OUP.

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

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

Gibson, A. and Asthana, S. (1998b) School performance, school effectiveness and the 1997 White Paper, *Oxford Review of Education*, 24, (2), 195-210.

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), 385-443.

Goldstein, H. (1998a) *Value Added pilot tables* http://www.ioe.ac.uk/hgoldstn/value-added-pilot-tables.htm

Goldstein, H. (1998b) Models for Reality A professorial lecture given at the London Institute of Education, July 1, 1998. http://www.ioe.ac.uk/hgoldstn/papers_for_downloading.htm

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, 50-53.

Kendall, L. (1995) Contextualisation of school examination results, 1992, *Educational Research*, 37, 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, 32-34.

Nuttall, D.L. and Willmott, A.S. (1972) *British Examinations: Techniques of Analysis*. Windsor: NFER Publishing Company.

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

Scottish Examinations Board (SEB) (1996) *Examination Statistics 1995*. Edinburgh: SEB.

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

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

SOED (1993) *Using examination results in school self-evaluation: VA indicators*, Section 8, Edinburgh: HMSO.

SOEID (1997a) *Examination Results in Scottish Schools 1994-96* Audit Unit, Edinburgh: HMSO.

SOEID (1997b) *How Good are our Results Audit Publications*, Edinburgh: HMSO.

SOEID (1997c) *Taking a closer look at attainment in secondary schools in Raising standards - setting targets*, Secondary Schools Support Pack, Edinburgh: HMSO.

SOEID (1997d) *The Improving Schools Effectiveness Project in Raising standards - setting targets*, Secondary Schools Support Pack, Edinburgh: HMSO.

Thomas, S and Mortimore, P. (1996) Comparison of VA models for secondary school effectiveness, *Research Papers in Education*, 11(1), 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: University of Strathclyde.

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

Tymms, P. (1996) Theories, models and simulations: school effectiveness at an impasse. in Gray, J., Reynolds, D., Fitz-Gibbon, C.T. and Jesson, D. (eds.) *Merging traditions: the future of research on school effectiveness and school improvement*. London: Cassell.

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.

Willms, J.D., (1992) *Monitoring school performance*. London: Falmer.