British Educational Research Journal, Vol. 27, No. 4, 2001

Predicting the Future: the role of past performance in determining trends in institutional effectiveness at A level

JOHN GRAY, *Faculty of Education, Cambridge University*
HARVEY GOLDSTEIN, *University of London Institute of Education*
SALLY THOMAS, *University of Bristol Graduate School of Education*

ABSTRACT *The analysis demonstrates that the average A/AS level results secured by English institutions from year to year are very stable. When account is taken of intake characteristics, however, correlations decrease substantially. The results show that predicting future 'value added' performance from past trends is unreliable. The implications of these findings for notions of 'continuous improvement' are discussed as well as their consequences for further research on institutional improvement.*

Introduction

> *You should remember that past performance is not necessarily a guide to future returns.* (Financial Services Authority)

The demand for schools to improve over time has become stronger in recent years. A mixture of external and internal forces has supported the pressures for change and improvement. Subtly but inexorably, a position has begun to emerge in which it is no longer sufficient simply to report that this year's results are at least as good as last year's—increasingly, there is an expectation that they will be better. Schools have responded, for their part, by setting themselves performance targets which have increased year on year.

A well-informed judgement that an institution's performance will rise the following year could be based on a number of factors. In committing themselves to increased figures, headteachers might, for example, have concluded that there had hitherto been a lack of will or effort on the part of those most closely involved but that this was now being remedied; that changes in personnel had occurred and that the new arrivals were

Received 25 October 2000; resubmitted 5 April 2001; accepted 12 April 2001.

ISSN 0141-1926 (print)/ISSN 1469-3518 (online)/01/040391-15 ©2001 British Educational Research Association
DOI: 10.1080/01411920120071425

more likely to deliver; that deficiencies in resources had been addressed; or that key staff members were now more knowledgeable about what to do than in the past. There again, they might simply have noted recent trends and extrapolated from them. Any or all of these factors might provide grounds for optimism/pessimism.

Earlier Research

Research on school effectiveness and school improvement should have something to contribute to these issues. However, in practice, comparatively little is known about trends in schools' performances over time. One reason for this has been because the necessary datasets for tracking institutional performance have not been available; another, that until recently researchers have been more interested in estimating the stability in schools' performances rather than changes. As Teddlie and colleagues (2000, p. 126) have observed, the research indicates that 'there is a fair degree of stability in secondary schools' effects on overall measures of achievement (e.g. total examination performance scores) over time (correlations are fairly strong and all positive)', a conclusion which is largely reinforced by Scheerens & Bosker's (1997) review of the international evidence. This conclusion corresponds with the commonsense notion that schools' performance does not vary much over time. However, the same review shows that most of the studies which have examined stability issues have only usually contained data on just two adjacent years. As Gray and colleagues have noted, data on three or more cohorts, along with some degree of *instability*, are necessary for schools to improve or decline in their effectiveness (Gray *et al.*, 1996). Unfortunately databases from which schools' 'improvement trajectories' could be constructed have only infrequently been available [1].

Earlier British studies have made some contribution to these issues. Using national databases on schools' examination results at the General Certificate of Secondary Education (GCSE), Gray and Jesson identified a number of schools which appeared to be improving over the period from 1992 through 1993 and 1994 to 1995 in a sustained year-on-year manner; they dubbed these schools 'consistent' improvers and judged that nationally about one in 10 of all secondary schools might fall into this category (see Gray *et al.*, 1999). A weakness of this study, however, was that no information on pupils' prior attainments at the time of their entry to secondary schools was available.

A study of all the secondary schools in one local education authority (LEA) by Gray *et al.* (1996) did take such changes in performance between intakes into account. Using data from five separate cohorts of pupils, it looked for schools which were clearly changing in their 'effectiveness' over time. A small minority of schools was indeed either improving or deteriorating over the 5 years. Evidence of additional schools on these sorts of 'trajectories' over 4 or 5 years was later found in further work undertaken in two other LEAs (Gray *et al.*, 1999).

Gray and colleagues concluded that evidence of *sustained* improvement was in relatively short supply: 'Up to three years of continuous improvement seems to have represented a good run for a school before it "reached a plateau" or "hit the wall" ' (see Gray *et al.*, 1999 p. 47). As they went on to argue, however, 'for a school to succeed in moving up a level of effectiveness required slightly more than this—four or five years of sustained improvement were necessary. In practice, very few of the schools (studied) managed to change groups in this way' (Gray *et al.*, p. 47).

In their separate ways, both these studies provide hints that predicting the future may be problematic. However, the absence of data on the prior attainments of their intakes,

in the former case, and the limited numbers of schools and LEAs involved, in the latter, restrict their utility for purposes of the present discussion.

A third and more recent study by Thomas (2001) has also examined trends over time using extensive datasets involving nearly 200 schools from London (over three cohorts) and Lancashire (over five cohorts). Using equivalent approaches and methods to those employed by Gray *et al.* (1996), the findings support and confirm, to some considerable extent, the evidence from previous research. However, it appears that the differences between London and Lancashire schools in value added time trends were smaller, and in some cases, not statistically significant.

There is a possibility that regional factors play a part in these variations in results across different datasets. For example, the extent of pupil selection within an LEA could influence the opportunities for a school to improve greatly their value added performance relative to other schools in the region (see tactics, discussed later). Whatever the case, Thomas (2001) emphasises that 'irrespective of many schools' apparent improvement in raw league table performance, only a small minority have been able to improve substantially in their effectiveness—*relative to that of other schools*' (original emphasis).

The study by Thomas (2001) also clarifies how different methods of analysis can result in different estimates of stability or instability in school effects over time. For example, using a separate analysis for each cohort (or separate intercept terms for each cohort in a joint analysis) emphasises instability over time with correlations of between 0.33 and 0.91 for cohorts differing by one year. These value added measures include both the stable and unstable elements of a school's effectiveness and represent schools' actual results for each individual cohort without the smoothing effect of fitting a trend. These measures demonstrate how short-term changes in policy and practice or random fluctuations over time (or both) can influence school performance from year to year. On the other hand, school effects that are modelled via a linear trend or 'rolling averages' (which combine the results of three consecutive cohorts) will accentuate stability over time, with correlations of between 0.77 and 0.99 for cohorts differing by one year. We shall extend this approach using the data in this article.

The analyses presented here build on the methodology of these earlier studies whilst the dataset itself has the advantage that it is nationally representative (see also Goldstein *et al.*, 2000). The age-group covered, however, relates to pupils proceeding from GCSE to A levels.

Key factors contributing to the explanation of variations in A level performance have included various measures of prior attainment at GCSE and gender (with males apparently making greater progress than females). The importance of these has emerged repeatedly in various studies. Two other factors have also features in previous work. Formal designations of institutional status frequently figure in public presentations of examination performance, although the extent of differences between institutions is usually considerably reduced when differences in intakes are taken into account. Working on the same dataset as the present authors, Yang & Woodhouse (2001) also report some evidence of 'compositional' effects—students in establishments where the average attainment of students in terms of prior performance at GCSE was high tended to make slightly better progress [2].

Key Questions

An apparently simple question lies at the heart of this analysis. If there is an embryonic trend emerging in a school's results, to what extent is one justified in simply extrapolating it? It is a well-established tenet of the 'effectiveness' research paradigm that

institutions differ in their levels of effectiveness. At the same time, they also seem to differ in their rates of improvement and their 'capacities to improve', although the research evidence for these conclusions is rather less extensive and a good deal more tentative. Furthermore, institutions are likely, at any one time, to differ both in their willingness and their ability to effect changes and improvements. To what extent, therefore, is there some kind of momentum to their improvement efforts?

The concern of the present analysis is to explore differences between institutions in terms of these improvement trajectories. A complicating factor in recent years, however, has been the extent to which there have been rising national trends in terms of key performance indicators. This has presented particular issues when interpreting results from the GCSE examinations usually taken at age 16 plus, where the greater majority of schools have been able to claim that their results have been improving on a year-by-year basis because of national trends. What Gray and colleagues (1999) have termed 'tactical' approaches (such as entering already high-attaining pupils for more examinations and concentrating on pupils on the borderline of key national indicators such as 5 A*–C grades) have paid dividends here without necessarily signifying any fundamental shift in the institution's effectiveness. At least by contrast with the changes which have occurred at GCSE, the A level system has been considerably more stable. The number of A/AS level points secured by the average candidate, for example, rose by only 7% over the period 1994–97. Furthermore, whilst there has been a tendency for candidates to be entered for more subjects (with increasing numbers being entered for more than the standard three), the substantial time commitments required to study for *any* A level have acted as a major constraint.

Whilst various 'tactical' approaches to raising grades have undoubtedly occurred amongst institutions preparing students for A/AS levels, their incidence has probably been more limited than at earlier stages of the educational system. The belief that A-level examinations provide some sort of 'gold standard' has possibly exerted a further pressure for stability. In a system where there is relatively little external 'turbulence' and where opportunities to initiate major innovations are relatively restricted, we should anticipate that systematic changes in performance will be limited.

The issues explored in the present article, therefore, build upon two related research traditions—the fairly well-established concern to estimate the extent of stability in institutions' effectiveness over time and the more recent interest in understanding changes in their effectiveness (improvement/deterioration). At the same time, however, it seeks to extend them with a view to establishing the extent to which *sustained* trends can be identified.

The Main Dataset and Derived Variables

The datasets employed for this analysis (which were provided by the Department for Education and Employment), were very large indeed. As they have been described in greater detail by Yang & Woodhouse (2001), we confine ourselves here to outlining their main characteristics. They covered four separate cohorts (the years 1994, 1995, 1996 and 1997). Each year included the full cohort of students who attained the age of 18 during the year, attended a recognised educational establishment in England and entered at least one examination at A/AS level. Students' results in General Studies (both at A level and GCSE) were excluded. Students were also excluded if grades or records were missing for either GCSE or A/AS level examinations; about 2% of cases fell into

this category. Students whose establishments were not identified reduced the total by a further 1% [3].

For the purposes of this study, only institutions where data were available for cohorts of students in each of the four years were suitable; excluding such schools resulted in a further reduction of the overall dataset [4]. Our analyses again suggested that the missing data could be treated as random. In the event, the main analyses reported in this article were conducted on just under 700,000 students attending just over 2500 institutions, making up some 96% of the original sample.

The main outcome variable used in these analyses was an overall measure of performance at A/AS level. This was based on the so-called UCAS Points Scale, widely used for screening entrants to universities in the UK. A score of 10 points is given to each grade A at A level, 8 to a B, 6 to a C, 4 to a D and 2 to an E grade, with no points being awarded for failures or ungraded results. At AS level, 5 points are given to an A grade, 4 to a B, 3 to a C, 2 to a D and 1 to an E. The totals are then summed to create an overall A/AS performance score for each student [5].

The average A/AS score for students in the study in 1994 was 13.9 points (with a standard deviation of 9.5 points). This rose in successive years to 14.9 points in 1997 (with an s.d. of 9.6 points). The prior performances of these candidates in their GCSE examinations also rose over the same period. In 1994, they were averaging 51.2 points per student; by 1997, this had risen to 53.7 points. Other key information about the sample is presented in Table Ia.

Just over two-fifths of the sample (44% in 1994) were in maintained comprehensive schools (see Table Ib). These institutions typically take students from a range of backgrounds and abilities from the ages of 11 to 18. Around a further tenth (11%) of the sample were in grant-maintained comprehensive schools whose constitutional status differed at that time from maintained schools in that they were independent of their LEAs. In total, just over half the sample were in comprehensive schools. A levels are typically a central focus of the provision such schools make for post-16 students. About one-fifth of the sample (20% in 1994) were in independent selective schools paid for privately by parents. Preparing students for A levels offering entry qualifications for higher education is the almost exclusive concern of such institutions. A further tenth of the sample (11%) were in further education colleges. The mission of these institutions is more wide ranging. Examinations leading to vocational qualifications are a major concern, with A levels frequently being a minority concern; students in these colleges quite often combine study for one or two A levels with the pursuit of other vocational qualifications [6]. The remainder of the sample was spread across a variety of other institutional types (see Table Ib).

The Main Analyses

The first step in the modelling process was to establish the pattern of relationships across years in institutions' A/AS level performance using multilevel modelling (Goldstein, 1995). Table IIa shows the results when gender, institutional type and a separate term for each year were fitted but not students' prior attainment at GCSE (Model 1); in the fixed part of the model, all the contrasts for institution type are with comprehensive maintained schools.

Table IIb also shows the pattern of correlations between these results across years. All these correlations are high, with those between adjacent years ranging from 0.93 (1995 and 1996) to 0.95 (1994 and 1995). The correlations across two years are almost as high

TABLE I. Descriptive statistics for each year cohort

(a) Individuals

	1994	1995	1996	1997
No. of students	164,177	167,535	175,162	189,706
% female students	53.1	53.2	53.1	53.6
Mean age of students (years)	18.5 (0.3)	18.5 (0.3)	18.5 (0.3)	18.5 (0.3)
Mean GCSE examinations taken/student	9.1 (0.9)	9.2 (1.1)	9.2 (1.0)	9.3 (1.0)
Mean GCSE total score/student	51.2 (9.6)	52.6 (10.6)	52.9 (10.1)	53.7 (10.3)
Mean GCSE average score/student	5.63 (0.86)	5.67 (0.83)	5.74 (0.79)	5.73 (0.80)
Mean A/AS score/student	13.9 (9.5)	14.4 (9.6)	14.8 (9.7)	14.9 (9.6)

Note: Standard deviations in parentheses.

(b) Institutions

	1994 (%)	1995 (%)	1996 (%)	1997 (%)
Maintained comprehensive	43.9	43.3	43.1	43.2
Maintained selective	2.6	2.6	2.5	2.5
Maintained modern	1.5	1.5	1.4	1.7
Grant-maintained comprehensive	10.8	11.2	11.6	12.0
Grant-maintained selective	3.2	3.2	3.3	3.3
Grant-maintained modern	0.7	0.7	0.8	0.8
Independent selective	19.7	19.6	19.3	18.7
Independent non-selective	2.1	2.3	2.6	2.4
Sixth form college	4.2	4.1	4.1	4.0
Further education college	10.5	10.5	10.6	10.7
Maintained unknown	0.5	0.6	0.6	0.6
Independent unknown	0.2	0.2	0.1	0.1

at 0.90 (1994 and 1996) and 0.92 (1995 and 1997), again supporting the view that institutions whose students perform well one year are very likely indeed to have students 2 years later who will perform at around the same levels. Indeed, even the three-year correlation (between 1994 and 1997) is high at 0.88, reinforcing the same view. In short, knowing students' overall performances one year seems to provide a very good prediction of the general level of their results in following years as well. This conclusion confirms more cursory glances at schools' performances—there are few surprises.

Model 2 is the same as Model 1 but adds in information about pupils' prior attainments at GCSE, including the numbers of GCSE subjects taken and the average grades secured in them, as well as further variables and terms based upon them (see Table II). Not surprisingly, prior attainment at GCSE turns out to have a considerable impact on explaining variations in performance. Many of the same factors seem to contribute as before, although the strength of some of these relationships is diminished. The average performance levels of other students entering the same institutions (the so-called 'compositional effect') also emerges as significant. Boys seem to make greater progress between GCSE and A/AS level; and there was also an interaction between gender and the average GCSE scores of those in the institution. At the same time, the amount of variation attributable to institutional effects was also reduced.

The effects of the introduction of GCSE scores into the equation can be seen most obviously in the reduced correlations between the institutions' residual scores over time (see Table IIb). The correlations between adjacent cohorts now vary around 0.75, between two years around 0.62 whilst across three years (1994 and 1997) they fall as low as 0.55.

We now explicitly fit a trend over time, which of course removes year-to-year random fluctuations in the results. In essence, Model 3 is a revised version of Model 1; it is virtually identical to Model 1 except for the fact that it uses linear and quadratic terms to summarise the year effects in order to make straightforward inferences about time trends and predictions (see Table IIIa).

Model 4 is the same as Model 3 but includes attainment at GCSE in addition as well as various other variables capturing the 'compositional' effects. The analysis suggests that the trends across years are effectively linear [7]. In the random part of the model, only the linear trend coefficient varied across institutions.

Predicting Performance

The evidence from both the unadjusted and adjusted models can now be used to predict performance. For this purpose, information about the first three years is used to predict the year 4 effect. The fitted model and estimated institution residuals (intercept and trend) from Model 3 are used to estimate an institution effect at year 4 and these are then plotted against the residuals estimated from Model 5, that is using year 4 only. The correlation between these is 0.94. When the same is done using Models 4 and 6, that is adjusting for GCSE results, the resulting 'value added' correlation drops to 0.51.

The equations for both the unadjusted and adjusted analyses for year 4 only (Models 5 and 6 respectively) are presented in Table IV.

For the unadjusted model, the correlation between the residual for year 4 predicted from the model describing the previous three years and the actual residual estimated from the year 4 data was very high indeed at 0.94. In substantive terms, this suggests that the schools' overall A/AS level results in 1997 did not vary very much from previous years; if they were high one year they were likely to be so the next. This

TABLE II. Multilevel model for A/AS level results over four years (standard errors in brackets)

(a)	Model 1—Unadjusted	Model 2—Adjusted
Fixed Coefficients		
Year cohort		
1994	− 0.370 (0.011)	− 0.722 (0.006)
1995	− 0.314 (0.011)	− 0.707 (0.006)
1996	− 0.268 (0.011)	− 0.707 (0.006)
1997	− 0.241 (0.010)	− 0.689 (0.006)
Prior attainment at GCSE		
No. taken (centred at 9)		0.063 (0.001)
Ave. GCSE score (centred at 9)		0.729 (0.005)
Ave. GCSE squared		0.108 (0.003)
Ave. GCSE cubed		− 0.085 (0.003)
Ave. GCSE to 4th power		− 0.017 (0.001)
No. taken cubed		0.003 (0.000)
No. taken squared		0.014 (0.001)
Fourth order polynomial I*		0.075 (0.002)
Cubic grafted polynomial		− 0.007 (0.000)
Fourth order polynomial II		0.001 (0.000)
Gender		
Female	0.080 (0.002)	− 0.021 (0.002)
Interaction gender/GCSE ave.		− 0.079 (0.002)
Institutional type		
Maintained selective	0.569 (0.042)	0.087 (0.020)
Maintained modern	− 0.505 (0.055)	− 0.107 (0.028)
Grant-maintained comprehensive	0.040 (0.022)	0.042 (0.010)
Grant-maintained selective	0.603 (0.038)	0.100 (0.018)
Grant-maintained modern	− 0.561 (0.079)	− 0.143 (0.041)
Independent selective	0.637 (0.018)	0.180 (0.009)
Independent non-selective	0.012 (0.047)	0.085 (0.025)
Sixth form college	0.181 (0.034)	0.102 (0.016)
Further education college	− 0.494 (0.023)	− 0.224 (0.011)
Unknown type	0.075 (0.077)	0.054 (0.038)
Residual variation		
Student level variance	0.760 (0.001)	0.407 (0.001)
− *2 log likelihood*	1798564.0	1364589.0

(b) Institution level correlation matrix (variances on diagonal; standard errors in brackets)
A: Unadjusted model

	1994	1995	1996	1997
1994	0.123 (0.001)			
1995	0.95	0.121 (0.004)		
1996	0.90	0.93	0.126 (0.004)	
1997	0.88	0.92	0.95	0.120 (0.004)

B: Adjusted

	1994	1995	1996	1997
1994	0.039 (0.001)			
1995	0.80	0.036 (0.001)		
1996	0.61	0.70	0.033 (0.001)	
1997	0.55	0.63	0.76	0.028 (0.001)

Notes: in the fixed part of the model, the contrasts for institutional type are all with maintained comprehensive schools. *The full details of the next three variables are as follows: fourth order grafted polynomial at GCSE score = 0; fourth order polynomial at number taken = 0; fourth order polynomial at GCSE score = − 2. See Goldstein (1995) and Yang & Woodhouse (2001) for further modelling details.

TABLE III. Unadjusted and adjusted A/AS level scores, including time trends (standard errors in brackets). Years 1–3 only

(a)	Model 3—Unadjusted	Model 4—Adjusted
Fixed coefficients		
Intercept (base year = 1994)	− 0.373 (0.011)	− 0.720 (0.006)
Year cohort		
Year trend (linear term)	0.065 (0.004)	0.008 (0.003)
Year term (quadratic term)	− 0.007 (0.001)	0.001 (0.001)
Prior attainment at GCSE		
No. taken (centred at 9)		0.064 (0.001)
Ave. GCSE score (centred at 9)		0.728 (0.005)
Ave. GCSE squared		0.108 (0.003)
Ave. GCSE cubed		− 0.085 (0.003)
Ave. GCSE to 4th power		− 0.017 (0.001)
No. taken cubed		0.003 (0.001)
No. taken squared		0.015 (0.001)
Fourth order polynomial I*		0.075 (0.002)
Cubic grafted polynomial		− 0.007 (0.001)
Fourth order polynomial II		0.001 (0.000)
Gender		
Female	0.082 (0.004)	− 0.025 (0.003)
Interaction gender/GCSE average		− 0.079 (0.002)
Institutional type		
Maintained selective	0.564 (0.042)	0.094 (0.020)
Maintained modern	− 0.503 (0.055)	− 0.108 (0.028)
Grant-maintained comprehensive	0.038 (0.022)	0.042 (0.011)
Grant-maintained selective	0.587 (0.038)	0.100 (0.018)
Grant-maintained modern	− 0.568 (0.079)	− 0.145 (0.041)
Independent selective	0.632 (0.018)	0.184 (0.009)
Independent non-selective	0.019 (0.047)	0.089 (0.025)
Sixth form college	0.193 (0.033)	0.101 (0.016)
Further education college	− 0.487 (0.022)	− 0.226 (0.011)
Unknown type	0.082 (0.076)	0.053 (0.038)
Residual variation		
Student level variance	0.761 (0.001)	0.410 (0.001)
− 2 log likelihood	1798699.0	1366046.0

(b) Covariance matrix between institutions
A: Unadjusted

	Intercept	Gender	Year (linear)
Intercept	0.133 (0.004)		
Gender	− 0.013 (0.001)	0.010 (0.0017)	
Year (linear)	− 0.005 (0.001)	− 0.000 (0.0003)	0.004 (0.0002)

B: Adjusted

	Intercept	Gender	Year (linear)
Intercept	0.041 (0.001)		
Gender	− 0.004 (0.007)	0.007 (0.0004)	
Year (linear)	− 0.007 (0.0004)	0.000 (0.0002)	0.004 (0.0002)

TABLE IV. Unadjusted and adjusted A/AS level scores (standard errors in brackets). Year 4 only

(a)	Model 5—Unadjusted	Model 6—Adjusted
Fixed coefficients		
Intercept	− 0.217 (0.011)	− 0.683 (0.007)
Prior attainment at GCSE		
No. taken (centred at 9)		0.069 (0.002)
Ave. GCSE score (centred at 9)		0.745 (0.009)
Ave. GCSE squared		0.107 (0.006)
Ave. GCSE cubed		− 0.092 (0.006)
Ave. GCSE to 4th power		− 0.020 (0.001)
No. taken cubed		0.004 (0.000)
No. taken squared		0.023 (0.002)
Fourth order polynomial I*		0.083 (0.004)
Cubic grafted polynomial		− 0.011 (0.001)
Fourth order polynomial II		0.001 (0.000)
Gender		
Female	0.010 (0.006)	− 0.021 (0.002)
Interaction gender/GCSE average		− 0.090 (0.004)
Institutional type		
Maintained selective	0.536 (0.042)	0.040 (0.022)
Maintained modern	− 0.487 (0.060)	− 0.085 (0.034)
Grant-maintained comprehensive	0.040 (0.022)	0.032 (0.012)
Grant-maintained selective	0.588 (0.038)	0.082 (0.020)
Grant-maintained modern	− 0.564 (0.089)	− 0.146 (0.050)
Independent selective	0.664 (0.019)	0.147 (0.011)
Independent non-selective	0.043 (0.057)	0.062 (0.033)
Sixth form college	0.132 (0.033)	0.088 (0.017)
Further education college	− 0.551 (0.024)	− 0.244 (0.013)
Unknown type	0.178 (0.086)	0.072 (0.046)
Residual variation		
Student level variance	0.745 (0.002)	0.393 (0.001)
− 2 log likelihood	488604.4	365845.8

(b) Covariance matrix between institutions
A: Unadjusted

	Intercept	Gender
Intercept	0.124 (0.004)	
Gender	− 0.018 (0.002)	0.015 (0.002)

B: Adjusted

	Intercept	Gender
Intercept	0.031 (0.001)	
Gender	− 0.007 (0.001)	0.009 (0.001)

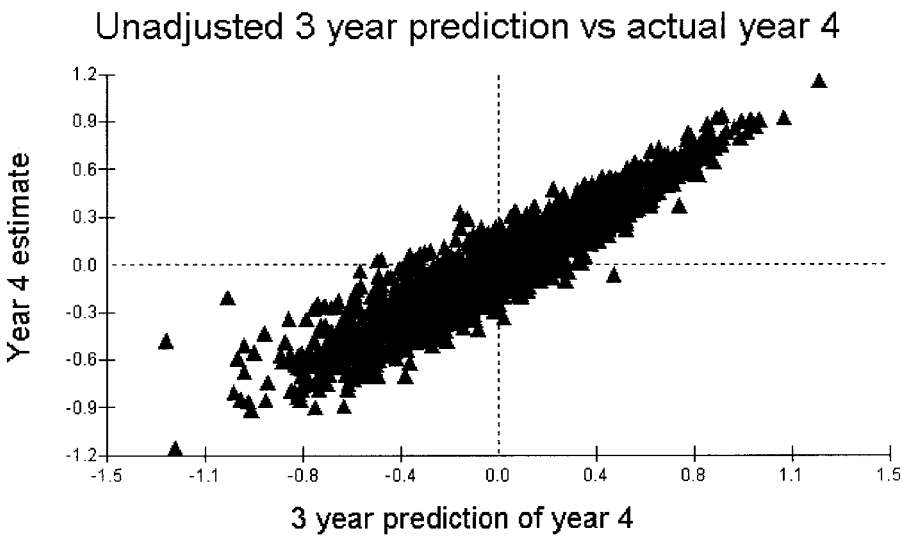

FIG. 1. Relationship between predicted year 4 results and actual year 4 results (Model 5—unadjusted).

relationship is displayed in Fig. 1. The evidence from the adjusted analyses, which take account of prior attainment at GCSE as well as compositional effects, provides a rather different picture. The correlation between the residual predicted from the model of trends over the previous three years and the actual year 4 residual was substantially lower at around 0.51 (see Fig. 2). In contrast to the picture emerging from the unadjusted

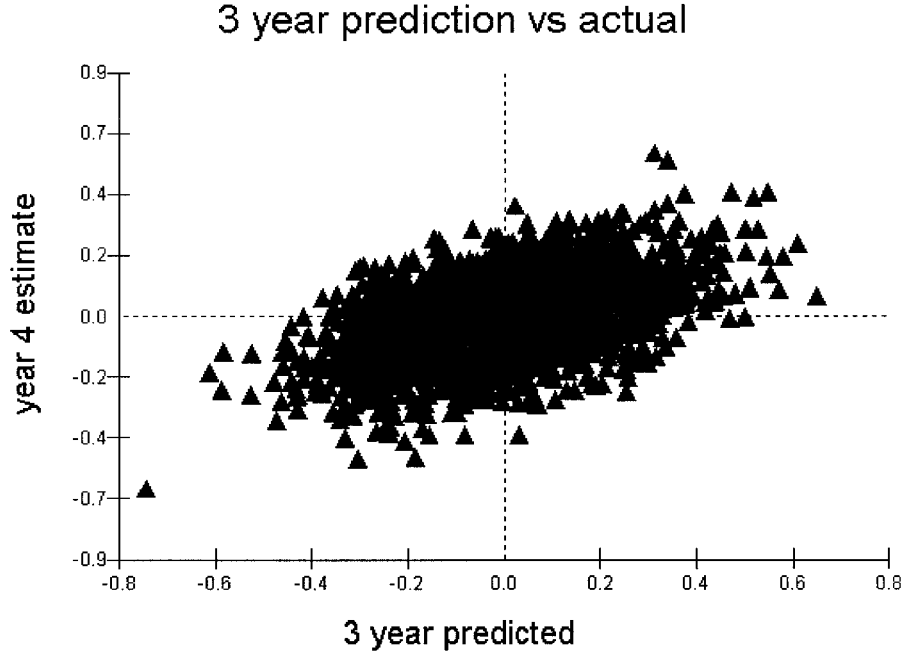

FIG. 2. Relationship between predicted year 4 results and actual year 4 results (Model 6—adjusted).

analyses, trends in 'effectiveness' over three years do not seem to provide a particularly reliable guide to 'effectiveness' with the fourth cohort. The correlation is moderate, demonstrating that past performance, at least using previous three-year trends, is not an accurate predictor of subsequent performance in the fourth year.

Using Trends to Predict Future Performance

Fig. 3 displays the time trend estimates across the four years for individual institutions in the study based on the adjusted model (Model 4) described earlier. For clarity of presentation, the graph is confined to a random 5% sample of all the institutions in the full analysis. [10] The plots also show the 95% confidence intervals surrounding each individual estimate.

In substantive terms, it should be noted that the trends in results over four years were fairly modest. The average rise in standardised point score, estimated from our model from 1994 to 1997, is around 0.15 in the unadjusted analysis. The most extreme institutions are estimated to have a rise of about 0.55 standardised points and a decrease of about 0.35 points. About 90% of institutions are estimated to have between a decrease of 0.05 and an increase of 0.35; this can be set against the between-student standard deviation of 0.87. In terms of A level points, this equates to a decrease of 0.5 points and an increase of 3.4 points, with a between-student standard deviation equivalent to 8.4 points.

For the adjusted analysis, the average rise from 1994 to 1997 in standardised point score was 0.04 points. The most extreme institutions ranged from a rise of about 0.11 points and a decrease of 0.03 points. About 90% of institutions are estimated to have somewhere between a decrease of 0.02 and an increase of 0.10. These changes need to be considered in the light of a between-student standard deviation of 0.64 and are

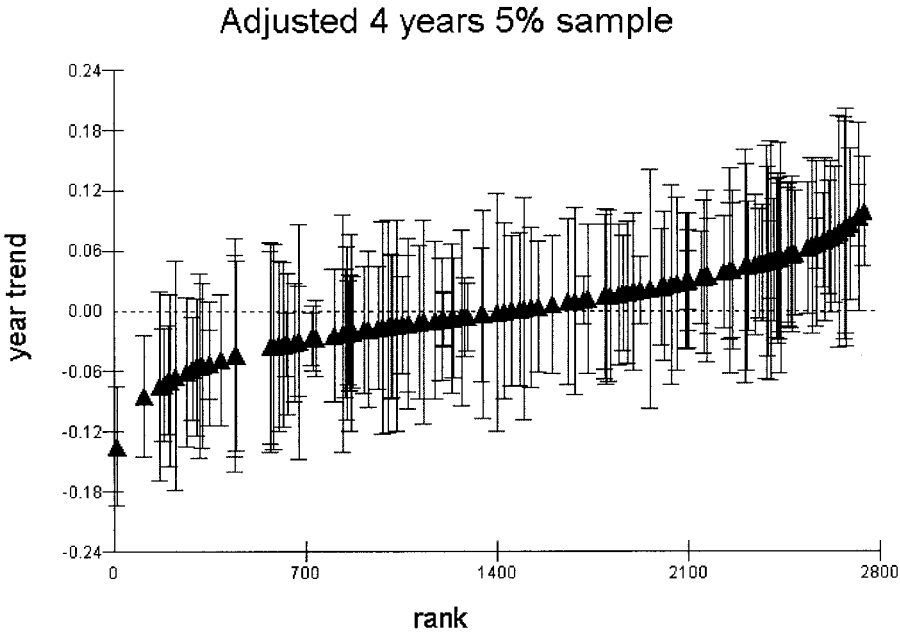

FIG. 3. Estimates of time trends (adjusted model—5% sample).

equivalent to an 0.2 point decrease and 1.0 point increase, with a between-student standard deviation of 6.1 points.

It is clear from Fig. 3 that the majority of these trajectories cannot reliably be distinguished from each other; only at the opposing ends of the distribution can institutions whose performance is clearly different be identified. Only a very small minority of institutions have improvement trajectories which can be described as sustained over time.

The implications of this finding for the *prediction* of future results are considerable. The trend is an estimate of the rate of change in A level scores each year. It is the slope term in the multilevel model. Two models were fitted to the data, one covering years 1–3 and the other years 3–4. The trend in years 1–3 was only mildly predictive of that in years 3–4; the correlations were 0.27 for the unadjusted trends and 0.28 for the adjusted trends (see Fig. 4).

Fig. 3 showed that the most marked and clearly distinguishable trends were confined to the extremes of the distribution. For this reason, we focus here on just the top and bottom 5% of institutions in years 1–3 whose trends were most marked over the years 1–3. Taking those institutions which were in the top 5% on the basis of their year 1–3 trends, these institutions were, on average, at about the twenty-fifth percentile of trends for years 3–4; this was true for both the adjusted and unadjusted models. Similarly, those institutions which were in the bottom 5% of trends for years 1–3 were at around the

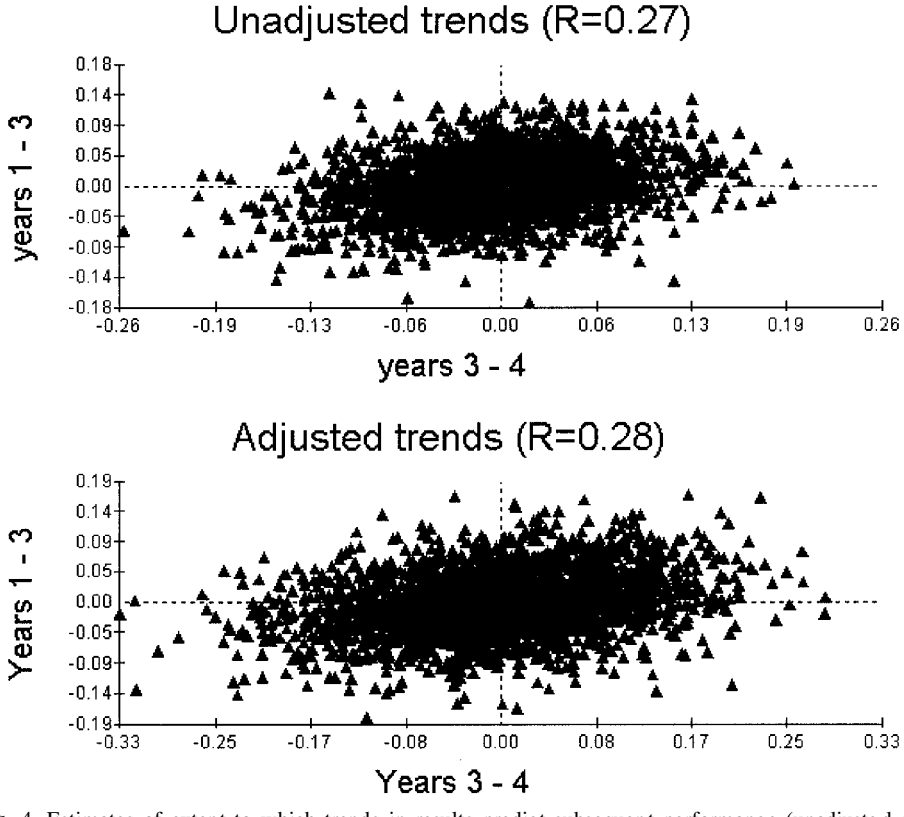

FIG. 4. Estimates of extent to which trends in results predict subsequent performance (unadjusted and adjusted models).

seventy-fifth percentile for years 3–4. Put another way, only just under one in seven (15%) institutions which were in the top (or bottom) 5% in terms of their trends for years 1–3 were still in the top (or bottom) 5% a year later for years 3–4. It needs to be borne in mind, however, that given the relatively modest extent of improvement and decline across the sample, such changes in performance were very modest in substantive terms.

Discussion and Conclusions

Several conclusions may be drawn from these analyses. First, the apparent levels of stability in institutions' results from one year to the next (correlations averaging 0.9 and above) are deceptive; they largely reflect the fact that most post-16 institutions attract similar sorts of students from one intake to the next. When students' prior attainments at GCSE are taken into account, the analyses suggest that there are substantial variations in effectiveness, as measured by value added models, over time. The levels of year to-year effectiveness discovered in this study are consistent with those reported in previous studies of institutional stability although they lie towards the lower end of the range of estimates produced to date.

Second, as in the very few previous studies which have explored issues relating to improvement, the evidence suggests that in the great majority of institutions preparing students for A levels, there are no obvious trends in effectiveness over time; only a very small minority of institutions have patterns which are consistently in an upward (or downward) direction.

Third, such trends as there are appear to be relatively short-lived and to come in bursts. In the present study, trends over 3 years were only mildly predictive of levels of effectiveness in the fourth year. There seem to be only a very small minority of institutions where there is some momentum to the trends over time. Indeed, the only confident prediction that could be made about a fourth year's results is that they were not very likely to reflect those which would be predicted by simply extrapolating from the trends. Among the small minority of institutions (the top and bottom 5%) where the trends were most evident over 3 years, the most likely outcome for the fourth year was in the opposite direction to the linear trend. This finding has hitherto only been demonstrated on much smaller local samples rather than nationally representative ones. Few institutions, it would seem, have yet managed to lock into cycles of 'continuous improvement'.

The urge to employ the past to predict the future is undoubtably strong. The present analysis provides some evidence of the likely limitations of such an approach. Whether there are other, more powerful strategies, however, remains a moot point.

Acknowledgements

This research was supported by the Economic and Social Research Council through a grant for the Analysis of Advanced Multilevel Modelling Methods for the Analysis of Examination Data (R000237394). The authors would like to thank the Department for Education and Employment for access to the datasets but stress that the analysis and interpretation are entirely their own responsibility.

Correspondence: John Gray, Homerton College, Hills Road, Cambridge CB2 2PH, UK.

NOTES

[1] The dominant emphasis in this article is on institutions whose performance improves over time; it should be noted, however, that the term 'improvement trajectory' also covers institutions whose performance deteriorates over time.

[2] In the analyses which follow, institutional type is routinely included as this is a public piece of information about the institutions. Compositional effects are ignored for similar reasons although there are arguments for including or excluding both variables. These models are not considered further here.

[3] There were no systematic differences between those in unidentified establishments and those in the rest of the dataset.

[4] 121 institutions had data for only three of the four years, 79 for two years and 60 for one year only. Inspection of the data on the establishments not represented in all four years revealed no dependence on establishment type, LEA or cohort size.

[5] Grades for General Studies and for multiple entries in the same subject have been excluded from the totals used in these analyses.

[6] The present study is confined to the A/AS level part of these students' studies.

[7] The quadratic term turns out to be non-significant but is retained in the analyses which follow.

[8] It would be possible to compare with fitting all four years and using the year 4 estimates from that but this would provide an essentially similar picture.

[9] The estimated 'true' year-on-year correlations for the adjusted residuals are likely to be higher than 0.5; this is because the estimated residuals contain sampling error and this 'noise' reduces the computed correlation.

[10] The graph for the unadjusted analysis has been omitted for reasons of space as it looks very similar to the adjusted analysis.

REFERENCES

GOLDSTEIN, H. (1995) *Multilevel Statistical Models* (London, Arnold).

GOLDSTEIN, H., HUIQI, P., RATH, T. & HILL, N. (2000) *The Use of Value Added Information in Judging School Performance* (London, Institute of Education).

GRAY, J., GOLDSTEIN, H. & JESSON, D. (1996) Changes and improvements in schools' effectiveness: trends over five years, *Research Papers in Education*, 11, pp. 35–51.

GRAY, J., HOPKINS, D., REYNOLDS, D., WILCOX, B., FARRELL, S. & JESSON, D. (1999) *Improving Schools: performance and potential* (Buckingham, Open University Press).

SCHEERENS, J. & BOSKER, R. (1997) *The Foundations of Educational Effectiveness* (Oxford, Pergamon).

TEDDLIE, C., REYNOLDS, D. & SAMMONS, P. (2000) The methodology and scientific properties of school effectiveness research, in: C. TEDDLIE & D. REYNOLDS (Eds) *The International Handbook of School Effectiveness Research* (London, Falmer Press).

THOMAS, S. (2001) Dimensions of secondary school effectiveness: comparisons across regions, *School Effectiveness and School Improvement*, 12(4).

YANG, M. & WOODHOUSE, G. (2001) Progress from GCSE to A and AS level: institutional and gender differences, and trends over time *British Educational Research Journal*, 27, pp. 245–267.