The limitations of using school league tables to inform school choice

by

George Leckie and Harvey Goldstein
University of Bristol, UK

Summary. In England, so-called 'league tables' based upon examination results and test scores are published annually, ostensibly to inform parental choice of secondary schools. A crucial limitation of these tables is that the most recent published information is based on the current performance of a cohort of pupils who entered secondary schools several years earlier, whereas for choosing a school it is the future performance of the current cohort that is of interest. We show that there is substantial uncertainty in predicting such future performance and that incorporating this uncertainty leads to a situation where only a handful of schools' future performances can be separated from both the overall mean and from one another with an acceptable degree of precision. This suggests that school league tables, including value-added ones, have very little to offer as guides to school choice.

Keywords: Examination results, Institutional comparisons, League tables, Multilevel modelling, Performance indicators, Ranking, School choice, School effectiveness, Value-added

1. Introduction

In many areas of the public sector, in the UK and elsewhere, performance indicators, in the form of rankings or 'league tables' are routinely published with the intention of helping to inform individuals' choice of institutions. For example, in health, waiting times and 'annual health check' scores are published to inform choice of NHS hospitals. In education, examination results are published to guide parental choice of schools for children about to enter each phase of schooling: primary schooling (ages 4-11), secondary schooling (ages 11-16) and a further two optional years of education (ages 16-18). A comprehensive review of the technical issues in these and other areas can be found in a report produced by the Royal Statistical Society (Bird et al., 2005). Common to these contexts is that the current performance of institutions is implicitly promoted as a guide to their future performance. However, no adjustment is made for the uncertainty that arises from predicting into the future. The present paper discusses this issue in the context of secondary school choice in England.

Secondary school league tables are published annually in England by the Department for Children, Schools and Families (DCSF) (formerly the Department for Education and Skills, DfES) (see: http://www.dcsf.gov.uk/performancetables). These tables are based on pupils' General Certificate of Secondary Education (GCSE) examination results taken at the end of compulsory schooling at age 16. These are important exams since successful GCSE results are often a requirement for progressing to studies for General Certificate of Education Advanced level (A-levels, ages 16-18) qualifications, themselves a common requirement for entry to university. For children who choose to leave education at age 16, their GCSE exam results are their only educational qualifications. The secondary school league tables allow inferences to be made about the performance of schools for the cohort of pupils who have just completed their secondary schooling (age 16). One of the principal aims of publishing these tables is to inform parental school choice for pupils who are just about to start secondary schooling (age 11). This was spelt out clearly by the Government led by John Major in the 'Parent's Charter' (DES, 1991) and has been endorsed by subsequent governments. The statistical uncertainty surrounding the estimated 'school effects' is typically expressed using 95% confidence intervals. However, there is additional uncertainty arising from the fact that these 'league tables' are always out of date since they refer to the performance of a cohort who began secondary schooling several years earlier (Goldstein and Leckie, 2008; Goldstein and Spiegelhalter, 1996; Wilson and Piebalga, 2008). Over this period, currently seven years, the performance of many schools changes considerably, limiting the extent to which current school performance can be used as a guide to future performance. Crucially, the league tables make no statistical adjustment for, nor do they warn about, the uncertainty that arises from predicting into the future.

In the education literature, value-added multilevel models are the preferred way of estimating school performance (for early examples see: Aitkin and Longford, 1986; Goldstein et al., 1993; Raudenbush and Bryk, 1986). These models adjust for pupils' intake achievement and other pupil characteristics known to affect educational achievement. These adjustments lead to a more relevant measure of a school's effect or contribution to the performance of its pupils than using simple school average GCSE scores. From 2006, the DCSF have used this methodology to estimate and publish a performance indicator for secondary schools that they term the 'contextual value-added' score (Ray, 2006).

The more school performance varies over time, the more misleading it will be to use current performance as a guide for parental choice. The literature on the stability of school effects has shown that, whilst measures of unadjusted achievement are highly correlated between cohorts, measures of value-added performance are far less so. These studies mostly consider

correlations for school effects over three, four and five years. At GCSE, Wilson and Piebalga (2008) report a correlation of 0.62 for value-added school effects two years apart whilst Gray et al., (1996) and Thomas (2001) report correlations of 0.56 and 0.55 respectively for four years apart. Similar magnitudes have been reported at A-level: Gray et al. (2001) and Yang and Woodhouse (2001) both report correlations of around 0.55 for school effects four years apart. The only study that has been able to look at correlations for longer than five years is Thomas et al. (2007) who, with 10 years of data, examine the stability of value-added school effects for schools within a single local education authority (LEA). They find a correlation of around 0.65 for school effects between cohorts five years apart and 0.62 for school effects between cohorts ten years apart. These correlations are higher than those reported in the other studies. One reason for this is that, unlike the other studies, these correlations are based on school effects that do not adjust for school level compositional variables, a point that we return to below.

In this paper, we use six cohorts of English data to show that there is indeed substantial uncertainty in predicting the future performance of schools. We present results using a multilevel model of school effectiveness that adjusts for prediction uncertainty. Section 2 provides background on school accountability and choice in England. Section 3 outlines the multilevel methodology. Section 4 describes the data and variables used in the analysis. Section 5 presents the main results; Section 6 presents our conclusions.

2. School league tables, accountability and choice

Since 1992, school rankings based on GCSE results have been published in school league tables. Initially, 'raw' performance measures such as the percentage of pupils gaining five or more GCSE passes at grade A*-C formed the basis for rankings. However, in 1995 the Government accepted the research evidence and agreed to move to a 'value-added' system whereby the prior achievements of pupils on entering secondary school would be used to make adjustments for different intake achievements resulting from explicit and implicit selection procedures. These simple value-added rankings were used between 2002 and 2005. Since 2006, so called 'contextual value-added' systems have been used which, in addition to adjusting for a pupil's own prior achievement, also attempt to adjust for factors such as the average prior achievement of a pupil's peers. Since 2006, the Government has also recognised that each school effect estimate should have an uncertainty (confidence) interval attached so that a statistically well informed judgement can be made about any differences between schools or differences between any one school and the population average. Thus, for example the DCSF web site (see: http://www.dcsf.gov.uk/performancetables) now provides intervals for contextual value-added estimates, although these are generally not prominent in media presentations or discussions. Similar performance indicators have been introduced to assess school performance over other stages of the education system: A-level, key stage 3 (KS3, ages 11-14 during secondary schooling) and key stage 2 (KS2, ages 7-11 during primary schooling).

The introduction of school league tables was originally justified on two distinct grounds, namely 'accountability' and 'school choice'. Concerns with the accountability of schools arose amid public debates about 'standards' and curriculum (see Goldstein, 2001 for further discussion of this). Holding schools publicly accountable for the performance of their pupils in GCSE examinations and later for their A-level and key stage test scores, was argued to be fair and would incentivise schools to improve their 'standards'. The rankings produced by this system are used by the national school inspection system of the Office for Standards in Education (Ofsted, http://www.ofsted.gov.uk/) and by Local Authorities with responsibilities for schools, to inform their judgements and also in some places as part of an accountability

screening system to identify those 'outlying' institutions that may require special attention (Yang et al., 1999).

For purposes of accountability, the most recent estimates of school effects are clearly the most appropriate ones to use, together possibly with estimates of trends over time. However, the situation with the second use of school rankings, for school choice, is rather different since the most relevant estimates would be for a future cohort. For example, for a cohort of 11 year olds entering schools in 2009, the relevant 16 year old exam results will be for the year 2014. However, when parents choose secondary schools (a year before their children enter secondary schooling), the most recently available 16 year old exam results are currently those published for the 2007 cohort. In other words, for the purpose of choice, what is required are *predicted* school effects some seven years beyond those typically currently available. It is these predictions that are explored in the present paper.

Before we describe the data and our analysis we need to consider carefully the basis for a useful prediction. Raudenbush and Willms (1995), and Willms and Raudenbush (1989) distinguish so called 'type A' school effects from 'type B' effects. The former are essentially those where adjustment has been made for a pupil's prior achievement and possibly other pupil characteristics. The latter effects additionally adjust for school 'compositional' factors such as the average prior achievement score or the average social composition of the pupil body. These variables measure the impact of pupils' peer groups on their achievement. Thus, type A effects are intended to inform parental school choice while type B effects are intended to assess those *practices* of the school that can be identified as responsible for school differences, that remain after controlling for school compositional variables. The distinction between type A and type B effects, however, is not always clear. Thus, schools may have some control over the social and intake composition of their pupils, linked for example to reputation, and it is not clear whether this should be adjusted for and whether it can really be separated from school practice.

From the point of view of school choice it seems clear that we should not adjust for any school level factors. The relevant question for a parent is whether, given the characteristics of their child, any particular school can be expected to produce better subsequent achievements than any other chosen school or schools. If a school level factor is associated with achievement this is strictly part of the effect being measured and not therefore something to be adjusted for. It is therefore Raudenbush and Willms's 'type A' effect that are essentially the ones we are considering. We note that the DCSF contextual value-added estimates do include school compositional effects and are therefore not appropriate for choice purposes. It is thus somewhat ironic that they have been promoted by government as improving choice. In the following exposition we shall not use any school compositional variables, although we will provide some comparisons with analyses that do use them.

3. Methodology

3.1 *Estimating school effects for the current cohort*

First we introduce the traditional school effectiveness multilevel model which provides value-added estimates of school performance for the current cohort. For simplicity, consider a two-level random intercepts model for pupils' GCSE scores where we treat pupils as nested within schools (for full details, see Goldstein, 2003). This model can be written as

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}, \qquad i = 1,\ldots,n_j, \qquad j = 1,\ldots,J$$
$$u_j \sim N\left(0,\sigma_u^2\right), \qquad e_{ij} \sim N\left(0,\sigma_e^2\right) \tag{1}$$

where y_{ij} is the GCSE score for the *i*th pupil within the *j*th school and x_{ij} is their prior achievement. u_j and e_{ij} are respectively the school level and pupil level random effects which are assumed normally distributed, independent of one another and independent of any predictor variables included in the model. Posterior or predicted estimates of the school effects u_j and their associated 'comparative' variances, that is $\text{var}\left(\hat{u}_j - u_j\right)$ which allow confidence intervals for the true values to be derived, are given by substituting sample estimates of the relevant parameters in

$$\hat{u}_j = \frac{n_j\sigma_u^2}{\left(n_j\sigma_u^2 + \sigma_e^2\right)}\tilde{y}_j, \qquad \text{var}\left(\hat{u}_j - u_j\right) = \frac{\sigma_u^2\sigma_e^2}{n_j\sigma_u^2 + \sigma_e^2} \tag{2}$$

where $\tilde{y}_j = 1/n_j \sum_{i=1}^{n_j}\left(y_{ij} - \beta_0 - \beta_1 x_{ij}\right)$ is the mean of the 'raw', fixed part, residuals for the *j*th school. The factor pre-multiplying $\tilde{y}_j$ is termed a 'shrinkage factor' since it moves the absolute value of the 'raw' mean residual towards zero. As the number of pupils in a school, n_j, increases the shrinkage factor tends to one and the variances tend to zero. Hence, school effects for large schools are shrunk less and estimated more precisely than those for smaller schools. Assuming normality, standard 95% confidence intervals for $\hat{u}_j$ are calculated as $\hat{u}_j \pm 1.96\sqrt{\text{var}\left(\hat{u}_j - u_j\right)}$. These shrunken school effects are published in the DCSF school league tables with 95% confidence intervals (see: http://www.dcsf.gov.uk/performancetables).

Typically it is supposed that parents are interested in making comparisons among a small set of schools. For any two schools, standard 95% confidence intervals are not appropriate for carrying out a significance test: they are too wide for this purpose. Goldstein and Healy (1995) propose an adjustment to these confidence intervals that makes two schools significantly different at the 5% level when their intervals just fail to overlap. For making a single pairwise comparison, they show that the width of these 'overlap intervals' should, on average, be approximately 1.4 times the standard error of the school effect in order to keep the overall significance level at approximately 5%. Note that this procedure is only appropriate for parents who make just one pairwise comparison; for comparing more than two schools a multiple comparisons procedure is required (see, for example, Afshartous and Wolf, 2007).

3.2 *Predicting school effects for future cohorts*

To make inferences about future cohorts of pupils, conditional on the currently observed cohort, we need to adjust both the estimates and the standard errors of the current school effects from model (1) to reflect the uncertainty that arises from predicting into the future.

Consider a bivariate response version of model (1) for two, not necessarily consecutive, cohorts of pupils (years of data)

$$\begin{aligned} y_{ij}^{(1)} &= \beta_0^{(1)} + \beta_1^{(1)} x_{ij}^{(1)} + u_j^{(1)} + e_{ij}^{(1)} \\ y_{ij}^{(2)} &= \beta_0^{(2)} + \beta_1^{(2)} x_{ij}^{(2)} + u_j^{(2)} + e_{ij}^{(2)} \\ \begin{bmatrix} u_j^{(1)} \\ u_j^{(2)} \end{bmatrix} &\sim N\left(0, \Omega_u\right), \qquad \Omega_u = \begin{bmatrix} \sigma_{u1}^2 & \\ \sigma_{u12} & \sigma_{u2}^2 \end{bmatrix} \\ \begin{bmatrix} e_{ij}^{(1)} \\ e_{ij}^{(2)} \end{bmatrix} &\sim N\left(0, \Omega_e\right), \qquad \Omega_e = \begin{bmatrix} \sigma_{e1}^2 & \\ 0 & \sigma_{e2}^2 \end{bmatrix} \end{aligned} \tag{3}$$

where superscripts '(1)' and '(2)' denote cohort 1 and cohort 2. Hence $y_{ij}^{(1)}$ is the GCSE score for the *i*th pupil in the *j*th school in cohort 1 whilst $y_{ij}^{(2)}$ is the GCSE score for the *i*th pupil in the *j*th school in cohort 2. The level 2 school residuals in general will be correlated. The level 1 residuals for the two responses are modelled as independent as a pupil can only belong to one cohort. Hence, this is a bivariate model where the bivariate structure is at level 2 rather than in the traditional multivariate multilevel model where it is at both levels.

We wish to estimate a set of school effects for cohort 2 when we only have data for cohort 1. To simplify matters we assume that the between school variance is constant across the two cohorts $\sigma_{u1}^2 = \sigma_{u2}^2 = \sigma_u^2$ (this assumption is readily tested, for example using a likelihood ratio test). This leads to a correlation of $\rho_{u12} = \sigma_{u12} / \sigma_u^2$ that is a measure of the stability of school effects between the two cohorts. It can then be shown (see Appendix A) that the posterior estimates and the associated comparative variances for a set of cohort 2 school effects based only on cohort 1 data are given by

$$\hat{u}_j^{(2)} = \frac{\rho_{u12} n_j^{(1)} \sigma_u^2}{\left(n_j^{(1)} \sigma_u^2 + \sigma_{e1}^2\right)} \tilde{y}_j^{(1)}, \qquad \operatorname{var}\left(\hat{u}_j^{(2)} - u_j^{(2)}\right) = \frac{n_j^{(1)} \sigma_u^4 \left(1 - \rho_{u12}^2\right) + \sigma_u^2 \sigma_{e1}^2}{n_j^{(1)} \sigma_u^2 + \sigma_{e1}^2} \tag{4}$$

where $\tilde{y}_j^{(1)}$ is the mean of the raw residuals for the *j*-th school in cohort 1. Comparing equation (4) with equation (2), we see that the posterior estimates for these 'future' school effects are smaller than the usual estimates, whilst their variances are larger. The shrinkage factor in equation (4) has an additional factor ρ_{u12} which further shrinks the future school effects towards zero but does not alter their rank ordering. As $\rho_{u12} \to 1$, the posterior estimates and associated variances for the future school effects will tend to the usual estimates for cohort 1. However, as $n_j^{(1)} \to \infty$ the future school effects tend to $\rho_{u12} \tilde{y}_j^{(1)}$ whilst the variances tend to $\sigma_u^2 (1 - \rho_{12}^2)$, rather than $\tilde{y}_j^{(1)}$ and zero respectively as is the case for the usual estimates for cohort 1. Hence, even for large schools, the estimates for future school effects will exhibit shrinkage and their standard 95% confidence intervals will be bounded.

Furthermore, the size of these effects will be expected to increase the further we predict into the future.

Thus, to estimate the future performance of schools we simply estimate model (1) for the current cohort of pupils and use equation (4) to obtain estimates and standard errors for the future school effects. We estimate model (3) on past data to obtain the correlation between school effects *t* cohorts apart. This makes the assumption that the *t* cohort apart correlation is stable over time, an assumption that is supported by the stability reported in the school effects literature.

We extend the models described here to include additional predictors measured at both levels. We restrict our analyses to random intercept models since these are widely used. However, we note that random coefficient models, which allow the coefficients of predictor variables to vary across schools, will often offer a more realistic description of the data (see, for example, Nuttall et al., 1989). More generally, the models can be extended to include, further levels, non-hierarchical data structures, discrete responses and multivariate responses (Goldstein, 2003; Raudenbush and Bryk, 2002). However, we do not pursue such extensions here although the same general results will be expected. All models are fitted using Iterative Generalised Least Squares (IGLS, Goldstein, 1986), that yields maximum likelihood estimates and is implemented in the MLwiN package (Rasbash et al. 2004).

4. Data

The exam data are taken from the national pupil database (NPD), a census of all pupils in the English state education system. The NPD holds data on pupils' test score histories and a limited number of pupil level characteristics. We extract six cohorts of pupils who took their GCSE or equivalent qualifications in 2002, 2003, 2004, 2005, 2006 and 2007. We include pupils with equivalent qualifications, such as those in vocational subjects, to be consistent with the published value-added school league tables. To these cohorts we match their KS2 exams taken five years earlier in 1997, 1998, 1999, 2000, 2001 and 2002 respectively. To each cohort, we then match data from the 2002 to 2007 pupil level annual school census (PLASC) datasets which contain data on pupil characteristics collected in the same year as their GCSE exams. (Further information on the NPD and PLASC datasets and how to access them can be found at http://www.bris.ac.uk/Depts/CMPO/PLUG/whatisplug.htm).

Our initial sample consists of 3373676 pupils spread over the six cohorts and nested within 3119 mainstream secondary schools. We exclude schools that have data for fewer than six cohorts and for convenience we exclude pupils who have missing values for any of the predictor variables used in the analysis. These exclusions reduce the sample to 2750430 pupils within 2657 schools and checks indicate that they can be regarded as a random subsample of the full dataset. To ease the computational burden of estimating our models, we choose to restrict the analysis to a 10 percent random sub-sample of the schools. This gives a final sample of 277583 pupils attending between them 266 schools.

4.1 *Variables used in the multilevel models*

As the response, we use a general attainment score that is the same as that used in the published value-added school league tables (for full details, see: http://www.dcsf.gov.uk/performancetables/nscoringsys.shtml). For each pupil, the general attainment score is defined as the sum of a pupil's individual scores across their separate GCSE and equivalent qualifications. The individual scores for each qualification are calculated using GCSE grades, recorded as: A* = 58, A = 52, B = 46, C = 40, D = 34, E = 28, F = 22, G = 16, U = 0. The general attainment score is then capped by choosing each pupil's

eight best grades. This measure is considered fairer than the total (uncapped) score since it lowers the scores of pupils who score highly merely by taking many examinations. The mean GCSE score is equivalent to eight grade C's, whilst a one standard deviation change is equivalent to a two grade change in each of the eight examinations. We treat the response as continuous and, so that the multilevel residuals better approximate the normality assumptions of the models, we monotonically transform the ranks of its values, within each cohort, to the corresponding expected values of order statistics from a standard normal distribution (Goldstein, 2003). Prior achievement is measured by pupil's KS2 average point score and is the same as that used in the published value-added school league tables. For each pupil, this is defined as their average score across their separate KS2 English, maths and science tests. To ease the interpretation of prior achievement in the analysis, the distribution of this variable is similarly transformed, within each cohort, to a standard normal score.

We choose to adjust for a similar set of pupil variables as those adjusted for in the published contextual value-added school league tables. These variables are gender, age within cohort (i.e. deviation in months from the mean age in the cohort), eligibility for free school meals (FSM, a proxy for low income), an indicator for special educational needs (SEN), an indicator for speaking English as an additional language (EAL), ethnicity and the income deprivation affecting children index (IDACI, a measure of residential neighbourhood social deprivation).

In our main analysis, we do not adjust for school-cohort compositional variables since we have argued these should not be included when the purpose of estimating the school effects is for school choice. However, since the DCSF do adjust for such variables we report how our results change when we also adjust for these variables. The compositional variables adjusted for are the mean and standard deviation of the intake achievement distribution for each school-cohort. These variables are constructed from the pupil level data and aim to capture the influence of pupils' peer groups.

4.2 *Description of the multilevel data structure*

The data consist of a three-level hierarchy of pupils nested within cohorts within schools. The median school-cohort has 190 pupils. Table 1 presents descriptive statistics for the six cohorts of pupils, 2002-2007. The percentage of pupils achieving five or more A*-C GCSE grades (5+A*-C) in 2002 is 55.2. This rose in successive years to 60.6 in 2007, with the largest increase occurring between 2005 and 2006. Over this period, the capped point score and prior achievement of these pupils also rose. The descriptive statistics for the pupil level characteristics suggest that they did not change markedly over the six years. Over the period as a whole, the percentage of FSM pupils has decreased slightly and there is some evidence that the percentage of Black African and in particular the percentage of pupils not belonging to one of the main ethnic groups have both increased.

Table 1. Means and standard deviations for pupils' background characteristics reported separately for each cohort: 2002-2007

Variable	2002	2003	2004	2005	2006	2007
Percentage of 5+A*-C GCSE pupils	55.2	55.9	56.6	58.5	60.7	60.6
Mean GCSE capped point score (original scale) [(i)]	296.8 (95.9)	292.8 (102.0)	294.6 (101.6)	298.7 (101.6)	304.4 (99.1)	308.3 (97.8)
Mean KS2 average point score (original scale) [(i)]	26.2 (4.1)	26.0 (4.0)	27.0 (4.1)	27.6 (4.0)	27.6 (3.9)	27.8 (4.0)
Percentage of female pupils	49.8	49.6	50.6	50.0	50.1	49.9
Percentage of FSM pupils	11.5	11.4	11.3	11.7	10.9	10.6
Percentage of SEN pupils	16.0	13.3	13.3	14.5	15.2	15.9
Percentage of EAL pupils	6.7	7.1	6.9	7.5	7.5	7.5
Ethnicity						
Percentage of White British pupils	87.1	87.2	87.1	86.3	86.2	86.4
Percentage of White non-British pupils	2.6	1.9	2.1	2.0	2.1	1.9
Percentage of Black Caribbean pupils	1.3	1.3	1.4	1.4	1.4	1.2
Percentage of Black African pupils	0.9	0.9	0.9	1.1	1.2	1.2
Percentage of Indian pupils	2.1	2.1	1.9	2.0	1.8	1.8
Percentage of Pakistani pupils	2.5	2.4	2.1	2.5	2.3	2.3
Percentage of Chinese pupils	0.3	0.3	0.3	0.4	0.3	0.3
Percentage of Other ethnic group pupils	3.2	3.8	4.1	4.4	4.8	4.9
Number of schools	266	266	266	266	266	266
Number of pupils	42949	44773	47229	46277	47851	48504

Note: (i) Standard deviations are reported in parentheses.

5. Results

5.1 *Estimating school effects for the current cohort*

Table 2 presents results from model (1), the traditional two-level random intercepts model of school effectiveness estimated on the 2007 cohort of pupils. The response is the normalised GCSE scores in 2007. The model adjusts for pupil prior achievement and the pupil background characteristics described in Section 4.1.

Table 2. Parameter estimates for the two-level random intercepts model of the normalised GCSE score for the 2007 cohort of pupils

	2007
Fixed Part	
Constant	-0.071 (0.014)
KS2 average point score	0.681 (0.005)
KS2 average point score (squared)	0.043 (0.003)
KS2 average point score (cubed)	-0.026 (0.001)
Female	0.184 (0.006)
Age within cohort (i)	-0.009 (0.001)
Free school meal (FSM)	-0.182 (0.010)
Special educational needs (SEN)	-0.373 (0.009)
English as an additional language (EAL)	0.326 (0.019)
Ethnicity (ref. White British)	
White non-British	0.096 (0.023)
Black Caribbean	0.071 (0.028)
Black African	0.194 (0.031)
Indian	0.143 (0.027)
Pakistani	0.026 (0.028)
Chinese	0.383 (0.057)
Other ethnic group	0.067 (0.016)
Neighbourhood social deprivation (IDACI) (ii)	-0.119 (0.004)
Random Part	
Between-school variance	0.046 (0.004)
Between-pupil-within-school variance	0.397 (0.003)
Deviance (-2*log likelihood)	93656
Number of schools	266
Number of pupils	48504

Note: Standard errors in parentheses. (i) The age within cohort variable ranges in values from -6 to +6 where -6 corresponds to the youngest pupil in the academic year (born on 31st August) and +6 corresponds to the oldest pupil in the academic year (born on 1st September). A one unit change in the variable corresponds to an age difference of one month. (ii) IDACI, is normalised to have mean zero, variance one.

In the fixed part of the model, KS2 average point score is entered as a cubic polynomial. This is necessary to adequately describe the non-linear graduation of the response to this variable particularly at the extremes. The effect of prior achievement is very strong; a one standard deviation increase in the KS2 average point score is associated with approximately two-thrids of a standard deviation increase in the GCSE score. The presence of the prior achievement measure effectively changes the interpretation of all subsequent variables in the model from explaining variation in achievement at GCSE, to explaining variation in progress made over secondary schooling. Girls and younger pupils make significantly greater progress than boys and older pupils. Those eligible for FSM and particularly those with SEN make significantly less progress whilst those speaking English as an additional language make more progress. All ethnic groups, particularly Black African, Indian and Chinese pupils, make considerably more progress than white-British pupils. Finally, those living in more deprived neighbourhoods make less progress.

The random part of the model separates the residual variation in GCSE scores into the parts that lie between-schools and within-school-between-pupils. The model gives a variance partition coefficient (VPC, Goldstein et al., 2002) of 0.104 $(= 0.046 / (0.046 + 0.397))$: 10.4% of unexplained differences in pupil progress are attributable to schools. Using equation (2) we estimate the school effect and associated standard error for each school in 2007. Fig. 1 plots these effects with 95% normal confidence intervals computed using 1.96 times the estimated standard errors so that this allows comparisons between each school and the average school.

Fig. 1 School effects for the 2007 cohort with 95% confidence intervals

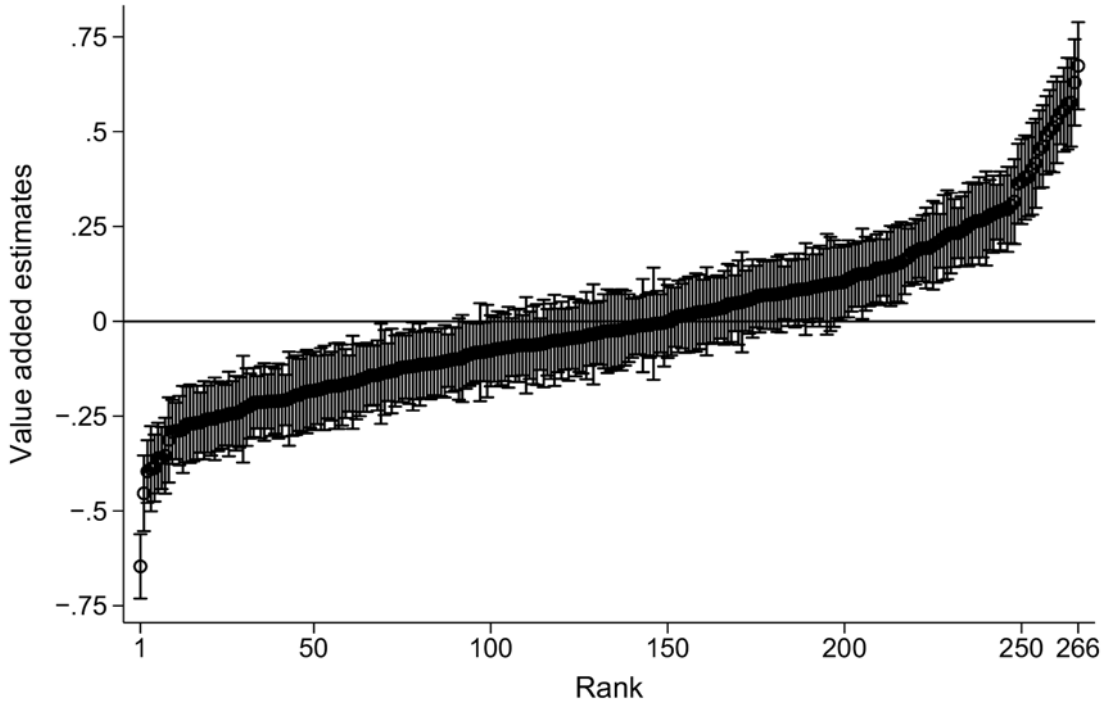
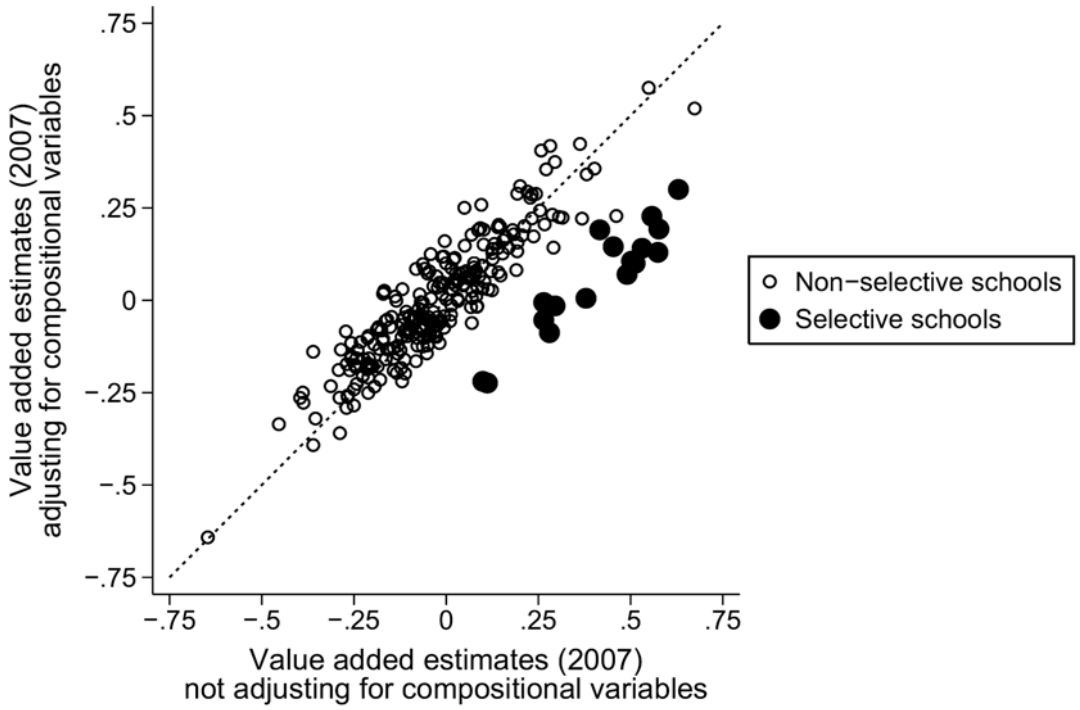


Fig. 1 illustrates the inherently imprecise nature of school effects (due to the small numbers of pupils within schools-cohorts); only 168 (63.2%) of schools are significantly different from the overall average. Importantly, this inference is only valid for the current cohort of pupils who sat their GCSE examinations in 2007. However, parents want to know whether the same significant differences will still apply for their children's cohort who will sit their GCSE examinations in seven years time, in 2014.

Adding the school-cohort level compositional variables (results not shown) suggests that there is a sizeable advantage in attending schools with a higher mean intake achievement, and to a lesser extent schools with a narrow spread as indicated by the standard deviation of intake achievement. Moving from the 10th to the 90th percentile in the school mean intake achievement distribution is associated with a 0.28 increase in a pupil's test score. The equivalent comparison for the spread of schools' intake achievements sees a decrease of 0.09 in a pupil's test score. In the random part of the model, adjusting for these variables halves the between-school variance, the VPC drops from 10.4% to 7.2% and now only 55.6% of schools are significantly different from the overall average. In sum, differences in the composition of schools' intake achievements, even after adjusting for pupils' own achievements, appears to be a major driver of between-school differences in GCSE scores.

Fig. 2 illustrates the association between the estimated school effects which do and do not adjust for compositional variables. The figure shows that adjusting for compositional variables substantially alters the school effects and rank positions for many schools, the correlation between the two sets of school effects is just 0.83. As we have already suggested, incorporating school level explanatory variables, for the purpose of parental school choice, is misleading since parents will want to know which school is best for their child, irrespective of whether this is due to school composition or due to school policies and practices. Importantly, the figure shows that the apparent performance of the 17 selective/grammar schools in one sample (indicated by large solid points) has worsened considerably when compositional effects are included, relative to non-selective schools (indicated by small hollow points). The selective admissions polices of grammar schools ensure that their pupils have a high mean and narrow spread of intake achievement. Hence, by including compositional variables we adjust for both peer group effects and a positive grammar school effect.

Fig. 2 Scatter plot of the 2007 school effects *adjusting* for school-cohort compositional variables against the 2007 school effects *not adjusting* for school-cohort compositional variables



5.2 *Predicting school effects for future cohorts*

In order to predict future school effects for 2014 given 2007 data, we need an estimate of the correlation between school effects seven years apart. However, we only have data for 2002-2007 which gives correlations between school effects for up to five years apart. Table 3 presents results from a bivariate response model based on the 2002 and 2007 cohorts of pupils. This model provides an estimate of the correlation between school effects five years apart and will therefore lead to a conservative picture of the inaccuracies that arise from predicting into the future. In the bivariate model, the two responses are pupils' normalised GCSE scores separately for each cohort. The model adjusts for prior achievement and the same set of pupil level variables as before.

Table 3. Parameter estimates for the bivariate two-level random intercepts model of the normalised GCSE score for the 2002 and 2007 cohorts of pupils

	2002	2007
Fixed part		
Constant	-0.055 (0.014)	-0.071 (0.014)
Average point score	0.667 (0.006)	0.680 (0.005)
Average point score (squared)	0.028 (0.003)	0.042 (0.003)
Average point score (cubed)	-0.026 (0.002)	-0.026 (0.001)
Female	0.189 (0.006)	0.184 (0.006)
Age within cohort [(i)]	-0.009 (0.001)	-0.009 (0.001)
Free school meal (FSM)	-0.217 (0.010)	-0.181 (0.010)
Special educational needs (SEN)	-0.412 (0.009)	-0.373 (0.009)
English as an additional language (EAL)	0.292 (0.021)	0.325 (0.019)
Ethnicity (reference is White British)		
White non-British	0.015 (0.025)	0.094 (0.023)
Black Caribbean	0.086 (0.028)	0.072 (0.028)
Black African	0.220 (0.035)	0.194 (0.030)
Indian	0.196 (0.027)	0.143 (0.027)
Pakistani	0.101 (0.028)	0.028 (0.028)
Chinese	0.237 (0.053)	0.383 (0.057)
Other ethnic group	0.162 (0.021)	0.067 (0.016)
Income deprivation affecting children index (IDACI) [(ii)]	-0.126 (0.004)	-0.117 (0.004)
Random Part: School [(iii)]		
Between-school variance (2002)	0.047 (0.004)	
Between-school variance (2007)	0.030 (0.004)	
Between-school covariance (2002, 2007)	0.047 (0.004)	
Random Part: Pupil		
Between-pupil-within-school variance (2002)	0.368 (0.003)	
Between-pupil-within-school variance (2007)	0.397 (0.003)	
Deviance (-2*log likelihood)	173243	
Number of schools	266	
Number of pupils	91453	

Note: Standard errors in parentheses. (i) The age within cohort variable ranges in values from -6 to +6 where -6 corresponds to the youngest pupil in the academic year (born on 31st August) and +6 corresponds to the oldest pupil in the academic year (born on 1st September). A one unit change in the variable corresponds to an age difference of one month. (ii) IDACI, is normalised to have mean zero, variance one. (iii) The school level variances have been restricted to equality.

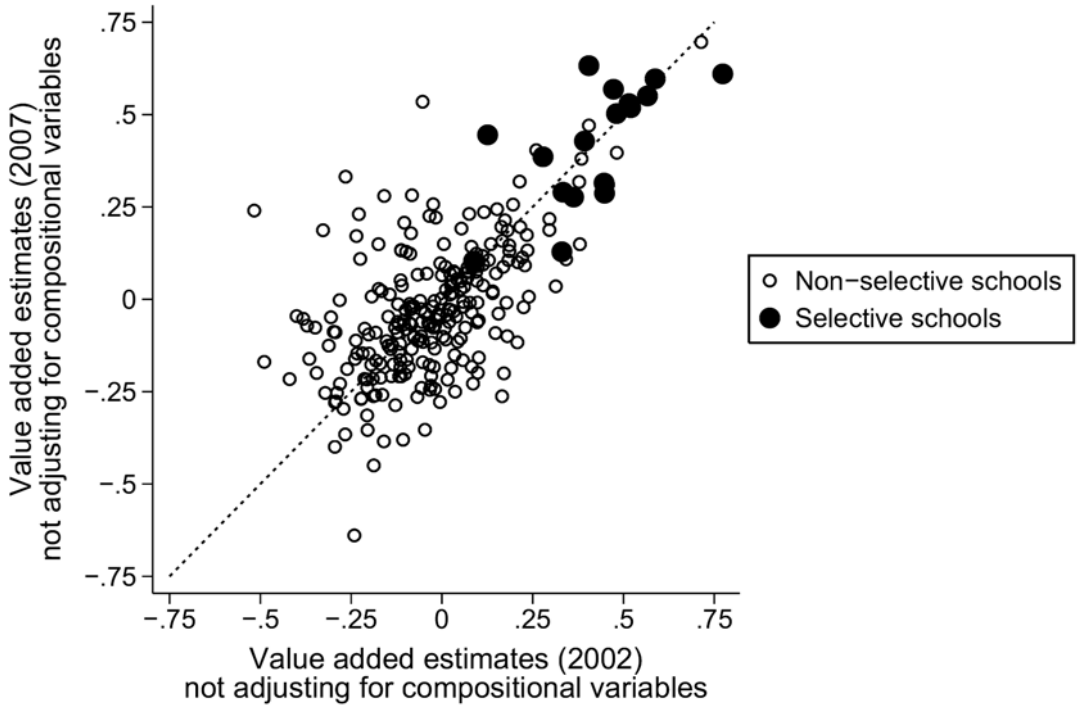
The fixed part parameter estimates for 2002 have the same signs and similar magnitudes to those for 2007. The magnitudes of the ethnic group estimates vary more across the two cohorts than is the case for the other predictor variables, but this is expected as these estimates tend to be estimated with (relatively) lower precision (due to the small number of pupils within certain ethnic groups). In the random part of the model, the between-school variances for the 2002 and 2007 cohorts are constrained to equal one another. A likelihood

ratio test shows that this constraint, which simplifies the formula for predicting future school effects, does not significantly reduce the fit of the model ($\chi^2_{(1)} = 0.156$, $p = 0.6929$). The model gives a VPC of 0.113 and 0.106 for 2002 and 2007 respectively; schools are no more or less important a source of variation in unexplained progress in 2007 than they are in 2002. The correlation between the 2002 and 2007 school effects, that will be used in the calculations for predicting the future school effects, is 0.64 $(= 0.030 / 0.047)$.

To show how the school effects become less stable over time, we estimate the bivariate model four more times, for 2007 and each of the 2006, 2005, 2004 and 2003 cohorts. The fixed and random part parameter estimates are fairly stable across all six cohorts (results not shown). However, the strength of the correlations between school effects decay the further apart the cohorts are; the correlations between 2007 and earlier cohorts are 0.89, 0.87, 0.76, 0.70 and 0.64. These correlations are higher than those reported in the literature, but this is expected since the literature has often adjusted for compositional variables whereas we do not. Indeed, if we do control for the school-cohort level compositional variables, the correlations between 2007 and earlier cohorts drop to 0.80, 0.73, 0.57, 0.46 and 0.40. Hence, the stability of school effects is in part due to compositional differences in schools' intakes that persist across schools over time.

Fig. 3 illustrates the association between the estimated school effects for the two cohorts furthest apart, 2002 and 2007. The figure shows that there are many schools with relatively high school effects in 2002 that have low school effects in 2007 and vice versa. Thus, using current school effects to make inferences five years into the future will result in many highly inaccurate judgements.

Fig. 3 Scatter plot of the 2007 school effects against the 2002 school effects



Equation (4) is used to predict estimates and standard errors for the future school effects in 2014. In these calculations we used the 2007 school effects and parameter estimates reported in Table 2 and the correlation of 0.64 for school effects five years apart from Table 3. Fig. 4

plots these effects with 95% confidence intervals that allow comparisons between each school and the average school.

Fig. 4 Future school effects for the 2014 cohort based on 2007 data with 95% confidence intervals

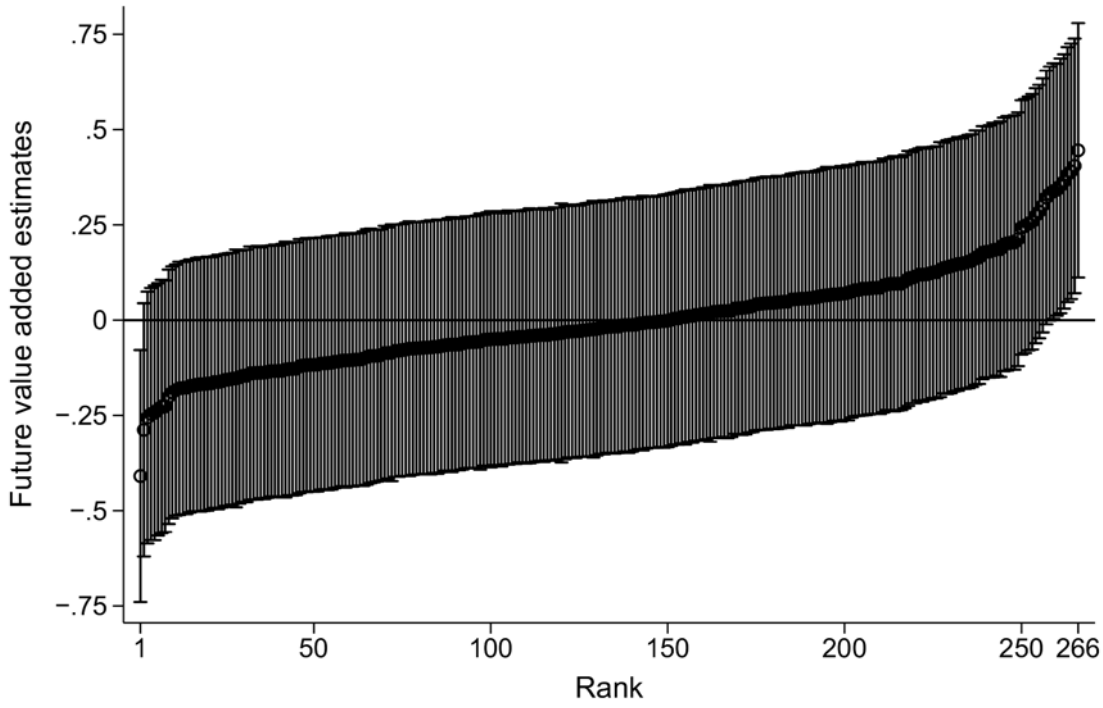


Fig. 4 shows that the predicted 2014 school effects have smaller magnitudes and wider confidence intervals than those for the 2007 cohort (see Fig. 1). The estimates of the future school effects have been adjusted towards zero and this reflects the fact that they contain less information about the likely 'effectiveness' of schools in 2014 than they do about the effectiveness of schools in 2007. In addition, their confidence intervals are widened to again reflect the increased statistical uncertainty involved when predicting into the future. On average, the confidence intervals are 3.5 times as wide as in Fig. 1. In Fig. 4 all but 9 (3.4%) of the 95 percent confidence intervals overlap zero. Hence, we can only predict, at the 5% level, that a handful of schools in 2014 will be significantly different from the average school. If we adjust for the school-cohort level composition variables, we find that no schools are significantly different from the overall average.

Many parents will not be interested in comparing a single school's performance to the average school; rather, they want to compare the performance of two schools with each other. Following the method proposed by Goldstein and Healy (1995), we construct an overlap interval for each school that is equal to the estimate of the school effect ±1.4 times its standard error. Using these overlap intervals, two schools are significantly different from each other, at the 5% level, if their overlap intervals fail to cross. We note that where parents wish to make more than one pairwise comparison, these overlap intervals should be wider (Afshartous and Wolf, 2007). Hence, the inferences we describe below give an optimistic picture of how well schools can be separated.

With 266 schools, the total number of possible pairwise comparisons is 35245 = 266(266 – 1)/2. For the 2007 cohort, we find that 62.7% of these allow significant separation. A similar percentage is found for the earlier cohorts. However, for the 2014 future school effects, the results show that only 2.1% allow significant separation and in the case when we adjust for

school-cohort compositional variables, no two schools can be separated. Another way of looking at this is if we use the 2002 data, then only 70.0% of 2007 significant separations are correctly identified. However, the 2002 data incorrectly identifies 6.4% (10.5%) of the 2007 pairwise comparisons where school *i* is significantly better (worse) than school *j* as school *i* being significantly *worse* (*better*) than school *j*.

In sum, having adjusted for the uncertainty of predicting five years into the future, we find that, at the 5% level, almost no schools are significantly different from the average school and very few schools can be predicted to be significantly different from each other at the 5% level. We also note that these are almost certainly upper limits since we have used the correlation appropriate to cohorts five years apart rather than seven years apart.

6. Conclusions

The purpose of this paper has been to demonstrate that, for purposes of school choice, using current school performance as a guide to future school performance is highly misleading. One justification for the publication of school league tables is that they are able to inform parental school choice. However, these tables do not adjust for prediction uncertainty, nor do they provide a clear statement of this statistical shortcoming. Importantly, when we account for prediction uncertainty, the comparison of schools becomes so imprecise that, at best, only a handful of schools can be significantly separated from the national average, or separated from any other school. This implies that publishing school league tables to inform parental school choice is a somewhat meaningless exercise. In addition, as we have pointed out, the current inclusion of compositional variables is inappropriate as the effects of these variables are part of the school effects that parents are interested in. See also Benton et al. (2003) who show that the inclusion of compositional variables changes the rank order of school effects. The current practice of adjusting for the school level mean and spread of intake achievement considerably worsens the performance of grammar schools, relative to non-selective schools and this has important policy implications.

Our method of predicting the future performance of schools is presented to illustrate the flaws with using the traditional school effectiveness model for choice purposes. It is not proposed as a new means of producing league tables. There are further reasons against using performance indicators to rank schools to inform choice since the statistical limitation discussed here is just one of a long list of concerns about using examination results as indicators of school performance (Goldstein and Spiegelhalter, 1996; Goldstein and Thomas, 1996). To the extent that parents may nevertheless wish to use the information that is provided by these tables, they will need to be aware that the uncertainty attached to them will necessitate a low weight being placed on them as compared with other sources of information available to parents through, for example, school inspections and local knowledge. However, we do feel that, used carefully, there is an *accountability* role for performance indicators as monitoring and screening devices to identify schools for further investigation, and, for this purpose, they should account for school composition and the most recent estimates are the most appropriate. For example, where these indicators find schools perform very well or very poorly for a cohort of pupils, it will often be interesting to study the policies and practices that these pupils were exposed to during their schooling. Nevertheless, for both monitoring and screening schools, performance indicators will be of most use when used together with other sources of school information; judgements based upon league tables alone should be considered as unsafe.

Our discussion has been in the context of parental choice of secondary schooling. However, our arguments and conclusions also apply to other phases of the education system such as

KS2 exams at 11 years and A-level exams at 18 years where, for the purpose of school choice, the most recent published data is currently six and four years out of date respectively. Our main result, that almost no schools can be significantly separated from each other, is likely to be even stronger for primary schools since these are, on average, a quarter the size of secondary schools. Finally, the statistical issues we discuss are also relevant to other public sectors such as health and social services, where attempts are also made to inform individual choices of institution based upon their past performance.

Appendix A

We consider here the special case of a level 2 repeated measures model where we just have two occasions at level 2, with the following structure. Using notation as in Goldstein (2003), Appendix 2.2.1, we write this model as

$$
\begin{aligned}
y_{ij}^{(1)} &= \left(X^{(1)}\beta^{(1)} \right)_{ij} + Z_i^{(1)} u_j^{(1)} + e_{ij}^{(1)} \\
y_{ij}^{(2)} &= \left(X^{(2)}\beta^{(2)} \right)_{ij} + Z_i^{(2)} u_j^{(2)} + e_{ij}^{(2)} \\
Z_i^{(k)} = \left\{ Z_{1i}^{(k)}, \ldots, Z_{pi}^{(k)} \right\}, &\qquad u_j^{(k)^T} = \left\{ u_{1j}^{(k)}, \ldots, u_{pj}^{(k)} \right\}, \qquad k = 1,2 \\
\begin{bmatrix} u_j^{(1)} \\ u_j^{(2)} \end{bmatrix} \sim N\left(0, \Omega_u\right), &\qquad \Omega_u = \begin{bmatrix} \Omega_{u11} & \\ \Omega_{u12} & \Omega_{u22} \end{bmatrix} \\
\begin{bmatrix} e_{ij}^{(1)} \\ e_{ij}^{(2)} \end{bmatrix} \sim N\left(0, \Omega_e\right), &\qquad \Omega_e = \begin{bmatrix} \Omega_{e11} & \\ 0 & \Omega_{e22} \end{bmatrix}
\end{aligned}
\tag{5}
$$

The superscripts '(1)' and '(2)' denote the cohorts. $X^{(k)}$ is the design matrix for the explanatory variables for cohort k, $\beta^{(k)}$ is the vector of covariate coefficients for cohort *k*, and $\left(X^{(k)}\beta^{(k)} \right)_{ij}$ is the linear fixed predictor for the *i*th pupil in the *j*th school for cohort *k*. The matrix $Z^{(k)}$ is the matrix of explanatory variables for the random coefficients which are assumed to have a joint multivariate normal distribution. The matrices Ω_u and Ω_e are the covariance matrices for the full set of level 2 and level 1 random coefficients respectively. The model is fitted to provide estimates of all the parameters and we consider estimating the set of level 2 residuals $\hat{u}_j^{(2)}$ given the observed data $y_{ij}^{(1)}$, $X^{(1)}$. Goldstein (2003, Appendix 2.2.1) provides expressions for the posterior estimates of residuals for a single occasion and their variances. Following that exposition and by considering the regression of the second occasion residuals on the first occasion raw residual estimates we obtain the required posterior or predicted residual estimates

$$
\hat{u}_j^{(2)} = E\left(u_j^{(2)} \middle| \tilde{Y}^{(1)}, \Omega_u, Z^{(1)} \right) = \Omega_{u21} Z^{(1)^T} V_1^{(1)} \tilde{Y}^{(1)} \tag{6}
$$

where $\tilde{Y}^{(1)} = Y^{(1)} - X^{(1)}\beta^{(1)}$ and their conditional or 'comparative' covariance matrix $\mathrm{cov}(\hat{u}_j^{(2)} - u_j^{(2)})$, which is used to provide interval estimates, is given by

$$
\mathrm{cov}(\hat{u}_j^{(2)} - u_j^{(2)}) = E\left\{ \left(\hat{u}_j^{(2)} - u_j^{(2)} \right) \left(\hat{u}_j^{(2)} - u_j^{(2)} \right)^T \right\} = \Omega_{u22} - \Omega_{u21} Z^{(1)^T} V_1^{(1)^{-1}} Z^{(1)} \Omega_{u12} \tag{7}
$$

where the final term in equation (7) is the 'diagnostic' covariance matrix $\mathrm{cov}(\hat{u}_j^{(2)})$. For completeness we note that equation (7) does not include the adjustment for the fact that the fixed part coefficients are estimated. These can be incorporated by replacing $V_1^{(1)^{-1}}$ by

$$V_1^{(1)^{-1}}\left\{V^{(1)}-X^{(1)}\left(X^{(1)^T}V_1^{(1)^{-1}}X^{(1)}\right)^{-1}X^{(1)^T}\right\}V_1^{(1)^{-1}} \qquad (8)$$

which provides a restricted maximum likelihood estimate (REML). Since the sample size in our data set is very large the REML adjustment is not needed. In this paper, we analyse a special case of the above where we fit a random intercept model at each occasion, that is $p=1$ and $Z_1^{(k)}$ is a vector of ones. By making the simplifying assumption $\sigma_{u1}^2=\sigma_{u2}^2=\sigma_u^2$ and writing $\rho_{u12}=\sigma_{u12}/\sigma_u^2$ we obtain the posterior estimates and the associated comparative variance for a set of cohort 2 school effects based only on cohort 1 data. These constraints applied to (6) and (7) lead to the expressions given in equation (4).

Acknowledgements

We are grateful to Simon Burgess, Stephanie von Hinke Kessler Scholder and Deborah Wilson for useful comments on the paper. The helpful comments from two referees and the journal editors are also gratefully acknowledged. Financial support from the Economic and Social Research Council (PTA-042-2005-00012) is gratefully acknowledged.

References

Afshartous, D. and Wolf, M. (2007) Avoiding 'data snooping' in multilevel and mixed effects models. *Journal of the Royal Statistical Society: Series A*, 170, 1035-1059.

Aitkin, M. and Longford, N. (1986) Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A*, 149, 1-43.

Benton, T., Hutchinson, D., Schagen, I. and Scott, E. (2003) Study of the Performance of Maintained Secondary Schools in England. National Foundation for Educational Research (NFER) (http://www.nfer.ac.uk/publications/other-publications/downloadable-reports/pdf_docs/NAO.pdf).

Bird, S. M., Sir David, C., Farewell, V. T., Harvey, G., Tim, H. and Peter C, S. (2005) Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society: Series A*, 168, 1-27.

DES (1991) The Parents Charter. London, Department of Education and Science

Goldstein, H. (1986) Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

Goldstein, H. (2001) Using Pupil Performance Data for Judging Schools and Teachers: scope and limitations. *British Educational Research Journal*, 27, 433-442.

Goldstein, H. (2003) *Multilevel statistical models 3rd Edition*, London, Arnold.

Goldstein, H., Browne, W. and Rasbash, J. (2002) Partitioning Variation in Multilevel Models. *Understanding Statistics*, 1, 223 - 231.

Goldstein, H. and Healy, M. J. R. (1995) The graphical presentation of a collection of means. *Journal of the Royal Statistical Society A*, 158, 175-177.

Goldstein, H. and Leckie, G. (2008) School league tables: what can they really tell us? *Significance*, 5**,** 67-69.

Goldstein, H., Rasbash, J., Yang, M., Woodhouse, G., Pan, H., Nuttall, D. and Thomas, S. (1993) A Multilevel Analysis of School Examination Results. *Oxford Review of Education*, 19, 425-433.

Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159, 385-443.

Goldstein, H. and Thomas, S. (1996) Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society: Series A*, 159, 149-163.

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

Gray, J., Goldstein, H. and Thomas, S. (2001) Predicting the Future: the role of past performance in determining trends in institutional effectiveness at A level. *British Educational Research Journal*, 27, 391-405.

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

Rasbash, J., Steele, F., Browne, W. and Prosser, B. (2004) *A User's Guide to MLwiN version 2.0*, London, Institute of Education.

Raudenbush, S. and Bryk, A. S. (1986) A Hierarchical Model for Studying School Effects. *Sociology of Education*, 59, 1-17.

Raudenbush, S. W. and Bryk, A. S. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods 2nd Edition*, Sage Publications Inc.

Raudenbush, S. W. and Willms, J. D. (1995) The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 307-335.

Ray, A. (2006) School Value Added Measures in England. Paper for the OECD Project on the Development of Value-Added Models in Education Systems. London, Department for Education and Skills (http://www.dcsf.gov.uk/research/data/uploadfiles/RW85.pdf).

Thomas, S. (2001) Dimensions of Secondary School Effectiveness: Comparative Analyses Across Regions. *School Effectiveness and School Improvement*, 12, 285-322.

Thomas, S. (2007) Modelling patterns of improvement over time: value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education,* 33, 261-295.

Willms, J. D. and Raudenbush, S. W. (1989) A Longitudinal Hierarchical Linear Model for Estimating School Effects and Their Stability. *Journal of Educational Measurement,* 26, 209-232.

Wilson, D. and Piebalga, A. (2008) Performance Measures, Ranking and Parental Choice: An Analysis of the English School League Tables. *International Public Management Journal,* 11, 344-366.

Yang, M., Goldstein, H., Rath, T. and Hill, N. (1999) The Use of Assessment Data for School Improvement Purposes. *Oxford Review of Education,* 25, 469-483.

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