England's Multilevel Model Based Value-Added School League Tables: Measuring and communicating statistical uncertainty to parents

George Leckie
Centre for Multilevel Modelling, University of Bristol, UK
E-mail: g.leckie@bristol.ac.uk

Abstract

In England, the Government annually publishes a range of school performance measures based on students' attainments in national assessment exercises. An important justification for their publication is that these measures should help parents make meaningful choices as to where they send their children to school. The most sophisticated, referred to as 'contextual value-added' (CVA), is derived from a multilevel model which adjusts for student intake differences between schools in order to more closely measure the effects schools have on their students. Schools' CVA performances are presented as point estimates with 95% confidence intervals to communicate their statistical uncertainty. However, for parents choosing schools, it is schools' future performances when their children will actually attend these schools which are most relevant, not schools' current performances. Another concern is that many parents find CVA estimates and especially their 95% confidence intervals, difficult to interpret. Indeed, when the media republish schools' CVA performances, they do so in the form of 'school league tables' whereby schools are ranked by their estimated performances and the 95% confidence intervals are omitted altogether. In this paper we extend the Government's multilevel model to demonstrate the additional uncertainty which arises when predicting schools' future performances. We then describe a simulation method to produce simple graphical summaries of the uncertainty in schools' predicted ranks as a more accessible alternative to presenting 95% confidence intervals.

Key Words: school league tables, multilevel model, statistical uncertainty, value-added model

1. Introduction

Each year, the English Government publishes school performance tables that report the attainment and progress of all secondary school (ages 11 to 16) students in the country. Their aim is to hold schools accountable, to encourage school self-evaluation and target setting, and to enable parents to make meaningful choices about where to send their children to school.

The first tables, introduced in the early 1990s, ranked schools based on simple school-level averages of students' performances in national examinations taken at the end of compulsory schooling (age 16). The tables were much criticized: not surprisingly, the highest scoring schools were largely those that recruited the highest attaining students at intake (age 11). The Government responded in 2002 by moved to a 'value-added' approach which adjusted for differences in students' intake scores between schools. However, such adjustments were not deemed sufficient to account for the differing socioeconomic contexts in which schools operate. Their approach was also descriptive and so failed to quantify the statistical uncertainty in measuring the effects schools have on their students. In 2006, the Government moved to a multilevel model (Goldstein, 2011) – also known as a mixed-effect model – based approach to estimating schools' performances which, as well as adjusting for students' intake scores adjusted for

students' socioeconomic backgrounds and reported the uncertainty in schools' estimated performances using 95% confidence intervals. The Government referred to these scores as 'contextual value-added' (CVA) scores. In 2011 the Government simplified the specification of their CVA model to no longer make adjustments for the differing socioeconomic contexts of schools' intakes arguing that doing so accepted that poorer students will do worse than richer students, rather than incentivising schools to narrow this achievement gap. The Government refer to this revised performance measure as value-added (VA). In this paper, we illustrate our arguments in terms of the old CVA measure, but the points we make apply equally to the newer VA measure.

In Goldstein and Leckie (2008) and Leckie and Goldstein (2009), we described a fundamental problem in using school performance tables, CVA or otherwise, for school choice: school league tables report the past performance of schools, based on children who have just taken their age 16 exams, whereas what parents want to know is how schools will perform in the future when their own children take their age 16 exams. Consider parents who chose a secondary school for their child in autumn 2011. Their child will enter school in autumn 2012 and will take their age 16 exams in 2017. Thus, the information parents need when choosing is how schools are predicted to perform in 2017. However, at the point of choosing, the most recent information was the school league table for how schools performed in 2010. There is therefore a seven year gap between the available information and what parents want to know. Clearly, the more schools' performances change over a seven year period, the less reliable school league tables will be as a guide to schools' future performances.

In Leckie and Goldstein (2011b), we described another concern with CVA which is many parents find the CVA estimates, especially their 95% confidence intervals, difficult to interpret. When the media republish schools' CVA performances, they attempt to simplify matters by publishing 'league tables' whereby schools are ranked by their estimated performances and 95% confidence intervals are omitted altogether. However, doing so implicitly encourages the public to interpret even the smallest differences in schools' league table positions as genuine differences on the quality of schools.

Table 1 illustrates the Government's presentation of schools' 2010 CVA performances for 19 schools in Bristol, one of 150 local authorities (approximately equivalent to a school district in the US) in England. The CVA scores are centred around 1000 and so schools scoring above 1000 are considered more 'effective' than the national average. Little attempt is made to explain the magnitude of these scores to parents, the aim appears to be for parents to implicitly rank schools and to simply examine which schools perform significantly above or below the national average. Of course, for parents choosing schools it is bespoke comparisons between several local schools which is of principal interest, not comparisons of each school in their local authority to the national average.

Table 2 reproduces the media's typical presentation of the same data. The 95% confidence intervals and school size are omitted from the presentation and the schools are explicitly ranked from the most effective to the least effective according to their CVA scores.

Table 1 The government's presentation of CVA

School	CVA score	Lower limit	Upper limit	Number of pupils
A	1003.2	993.7	1012.8	176
B	1001.0	990.7	1011.4	150
C	970.3	961.5	979.1	210
D	1004.7	994.7	1014.7	160
E	988.5	979.3	997.8	187
F	1003.4	992.8	1013.9	142
G	1015.7	1005.8	1025.5	164
H	1004.2	994.4	1013.9	169
I	1006.0	996.2	1015.8	167
J	1013.4	1004.4	1022.3	203
K	979.5	969.1	989.9	147
L	1036.2	1026.6	1045.8	174
M	1005.0	994.2	1015.8	134
N	1002.0	987.0	1017.0	65
O	1009.6	992.4	1026.7	47
P	1009.6	998.1	1021.2	117
Q	1007.4	995.4	1019.3	109
R	1028.2	1015.0	1041.4	87
S	1010.8	998.5	1023.1	102

Table 2 The media's presentation of CVA

Rank	School	CVA Score
1	L	1036.2
2	R	1028.2
3	G	1015.7
4	J	1013.4
5	S	1010.8
6	O	1009.6
7	P	1009.6
8	Q	1007.4
9	I	1006.0
10	M	1005.0
11	D	1004.7
12	H	1004.2
13	F	1003.4
14	A	1003.2
15	N	1002.0
16	B	1001.0
17	E	988.5
18	K	979.5
19	C	970.3

2. The Government's CVA multilevel model

The Government's multilevel CVA model is a two-level students (level-1) within schools (level-2) random-intercept multilevel model. Let y_{ij} and x_{ij} denote the age 16 and age 11 achievement scores for student i $(i = 1,2,\dots,n_j)$ in school j $(j = 1,2,\dots,J)$. The Government's model, expressed for simplicity in terms of only these two variables, can then be written as

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}, \\ u_j &\sim \mathrm{N}(0, \sigma_u^2), \\ e_{ij} &\sim \mathrm{N}(0, \sigma_e^2), \end{aligned} \tag{1}$$

where and u_j and e_{ij} are the school- and student-level random effects, assumed normally distributed with zero means and constant variances. The full multilevel model specification and parameter estimates for the Government's model can be found on their website.

The published CVA scores are calculated as the posterior (empirical Bayes) estimates of the school effects

$$\hat{u}_j = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_e^2}{n_j}} \frac{\sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})}{n_j},$$

where $\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij}$. Their associated 'comparative' variances, used to calculate the published 95% confidence intervals, are given by

$$\mathrm{var}(\hat{u}_j - u_j) = \frac{\sigma_u^2 \sigma_e^2}{n_j \sigma_u^2 + \sigma_e^2}.$$

For the purposes of school choice we wish to predict a set of CVA scores for the 2017 cohort based only on the 2010 cohort. Leckie and Goldstein (2009) proposed the following two-cohort version of the Government's model for this purpose

$$\begin{aligned}
y_{ij}^{(0)} &= \beta_0^{(0)} + \beta_1^{(0)} x_{ij}^{(0)} + u_j^{(0)} + e_{ij}^{(0)} \\
y_{ij}^{(7)} &= \beta_0^{(7)} + \beta_1^{(7)} x_{ij}^{(7)} + u_j^{(7)} + e_{ij}^{(7)} \\
\begin{pmatrix} u_j^{(0)} \\ u_j^{(7)} \end{pmatrix} &\sim \mathrm{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u(0)}^2 & \\ \sigma_{u(0,7)} & \sigma_{u(7)}^2 \end{pmatrix} \right\} \\
\begin{pmatrix} e_{ij}^{(0)} \\ e_{ij}^{(7)} \end{pmatrix} &\sim \mathrm{N}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{e(0)}^2 & \\ 0 & \sigma_{e(7)}^2 \end{pmatrix} \right\}
\end{aligned} \tag{2}$$

where the '(0)' and '(7)' subscripts and superscripts denote the 2010 and 2017 cohorts, respectively. By making the simplifying assumption that $\sigma_{u(7)}^2 = \sigma_{u(0)}^2$, it can be shown that the posterior estimates for the 2017 CVA scores based only on the 2010 data are given by

$$\hat{u}_j^{(7)} = \rho_{u(0,7)} \hat{u}_j^{(0)}$$

with associated variances given by

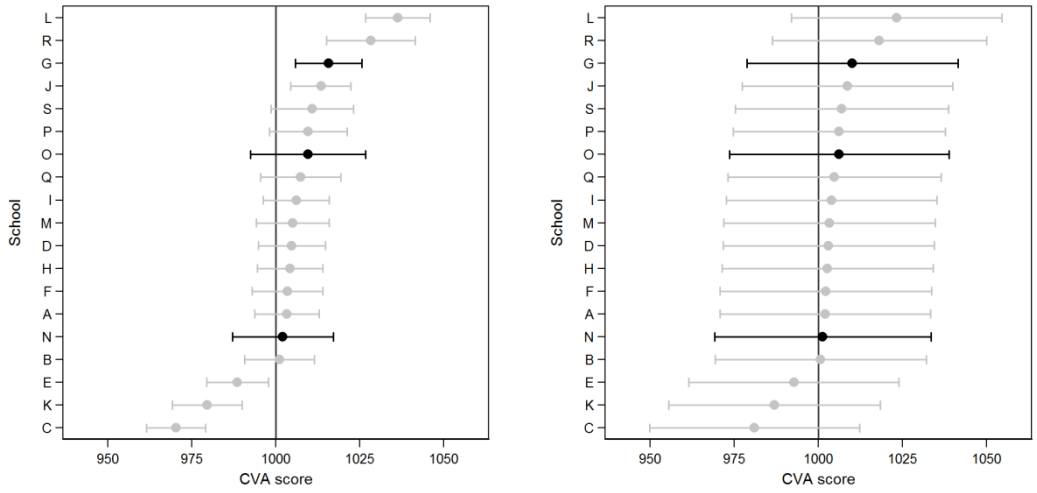
$$\mathrm{var}\left(\hat{u}_j^{(7)} - u_j^{(7)}\right) = \mathrm{var}\left(\hat{u}_j^{(0)} - u_j^{(0)}\right) + \frac{n_j^{(0)} \sigma_{u(0)}^4 \left(1 - \rho_{u(0,7)}^2\right)}{n_j^{(0)} \sigma_{u(0)}^2 + \sigma_{e(0)}^2}$$

The government publishes sample estimates for $\sigma_{u(0)}^2$ and $\sigma_{e(0)}^2$ and so the only unknown parameter is $\rho_{u(0,7)}$, the seven-cohort apart correlation between the 2010 and 2017 CVA scores. We substitute a value of 0.64 which is the correlation reported in Leckie and Goldstein (2009) from fitting (2) to two observed cohorts of data five years apart. The corresponding 95% confidence intervals for the CVA scores are calculated in the usual way. Leckie and Goldstein (2011a) generalize the approach described here to the case of predicting CVA scores for a future cohort when we observe multiple past cohorts, but we do not consider this extension further here.

Figure 1 plots the CVA estimates with 95% confidence intervals for the 2010 (left panel) and 2017 (right panel) cohorts. The 2010 estimates are the relevant estimates for holding schools accountable as they relate to how schools have performed for their current students. The 2017 estimates are the relevant estimates for school choice as they relate to how schools are predicted to perform for their 2017 students. In both panels, the schools

are ranked in the same order as they are based on the same 2010 data. However, the intervals for the 2017 cohort are substantially wider than those for the 2010 cohort reflecting the greater uncertainty associated with predicting schools' future performances compared to their current ones. Indeed, schools' future performances are so uncertain that no schools can be distinguished from one another with an acceptable degree of precision.

Figure 1 CVA estimates with 95% confidence intervals for the 2010 (left panel) and 2017 (right panel) cohorts.



3. Simulation method

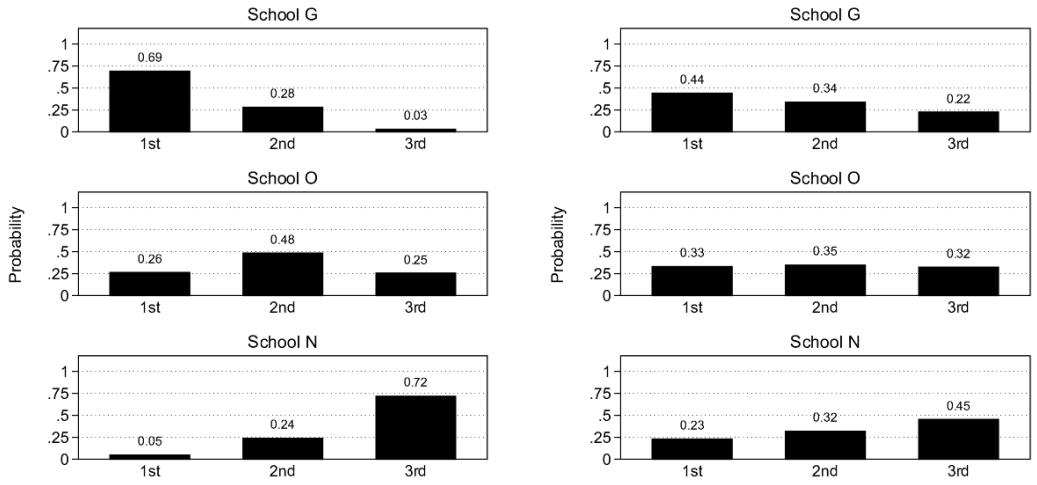
In this section, we describe a simulation method for making bespoke comparisons between several local schools and for producing simple graphical summaries of the uncertainty in the rank order of these schools. The method follows that presented by Goldstein and Spiegelhalter (1996) and Marshall and Spiegelhalter (1998) in their analysis of healthcare league tables. To illustrate the method, we focus on the three schools closest to the author's university work address: Schools G, N and O. Figure 1 suggests School G to be the best school, then School O and then School N. However, the confidence intervals for these three schools substantially overlap making this rank ordering far from certain.

To apply the simulation method, we appeal to a Bayesian interpretation of the sampling distributions of these three CVA scores whereby each distribution provides the probability distribution of each possible CVA score for that school. We repeat the simulation method 10,000 times where on each iteration we sample a value from each of the three schools' sampling distributions and then rank these values. We then calculate the proportion of the 10,000 iterations that each school was ranked 1st, 2nd or 3rd and interpret these probabilities as the probability that each school is the best, second best or third best school, respectively. Figure 2 reports these probabilities for the 2010 and 2017 cohorts.

Figure 2 clearly shows that while school G appears to be the best school, this is far from certain. The predicted probability that school G is the best school is just 0.69, compared to a probability of 0.26 that school O is the best school and a probability of 0.05 that school N is the best school. Furthermore, when we repeat the simulation for the 2017

cohort, the probability that school G, N or O is the best school reduces to 0.44, 0.23 and 0.33, respectively. Thus, we are far less certain that school G will be the best school for the 2017 cohort than we were that school G was the best school for the 2010 cohort.

Figure 2 Probabilities that school G, N and O are ranked 1st, 2nd or 3rd for the 2010 (left panel) and 2017 (right panel) cohorts



4. Conclusion

In this paper we have argued that for school choice purposes it is schools' future performances which are most relevant, not their current performances. We have shown that schools' 2010 CVA scores do not contain enough information to be able to make accurate predictions about schools' performances in seven years' time. We have illustrated this first in terms of the standard presentation of CVA scores in the form of estimates and 95% confidence intervals. We then presented a simulation method that enables schools to be ranked, but in a way which communicates the uncertainty in making such rankings through simple statements about the chance that each school has the highest score rather than through the current use of confidence intervals.

References

Goldstein, H. (2011). *Multilevel statistical models*. Wiley.

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

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

Leckie, G. & Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*, 835-851.

Leckie, G. & Goldstein, H. (2011a). A note on "The limitations of using school league tables to inform school choice". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174, 833-836.

Leckie, G. and Goldstein, H. (2011b). Understanding uncertainty in school league tables. *Fiscal Studies*, 32, 207-224.

Marshall, E. C. and Spiegelhalter, D. J. (1998). Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ*, 316, 1701-1704.