The limitations of using school league tables to inform school choice

George Leckie and Harvey Goldstein
Centre for Multilevel Modelling
University of Bristol
Introduction

• School league tables (i.e. school report cards) rank schools by:
  – Schools’ average test scores
  – Estimates of school quality based on statistical models

• They are published:
  – To hold schools accountable
  – To inform parental school choice

• They are now published in many countries
  – Australia, Canada, England, US,...
The English education system

• Two phases
  – Primary schooling from ages 5 to 11
  – Secondary schooling from ages 11 to 16

• Two main tests/exams
  – At age 11 children take English and maths tests
  – At age 16 children take GCSE exams
A brief history of England’s school league tables

• **1994 onwards:** Schools’ averages GCSE exam results
  – Unfair since schools differ in the quality of their intakes
  – Not model based, so no statement of statistical uncertainty

• **2006 onwards:** Contextual value-added (CVA) scores
  – Adjusts for the intake achievement of students
  – Based on a multilevel model, so scores are published with 95% confidence intervals
## Local Authority: Oxfordshire

<table>
<thead>
<tr>
<th>School</th>
<th>Cohort</th>
<th>KS4 Results</th>
<th>Key Stage 2 to 4 CVA</th>
<th>Expected progress from KS2-KS4</th>
<th>Year on year comparisons</th>
<th>Absence</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>% of pupils achieving</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Level 2 (5+ A*-C) A* C including English and maths GCSEs</td>
<td>English and maths Skills at Level 2</td>
<td>English and maths Skills at Level 1</td>
<td>Level 2 (5+ A*-C)</td>
<td>Level 1 (5+ A*-G)</td>
</tr>
<tr>
<td>Banbury School - Banbury</td>
<td>40%</td>
<td>47%</td>
<td>94%</td>
<td>68%</td>
<td>91%</td>
<td>82%</td>
<td>14%</td>
</tr>
<tr>
<td>Bartholomew School - Witney</td>
<td>66%</td>
<td>70%</td>
<td>99%</td>
<td>71%</td>
<td>99%</td>
<td>64%</td>
<td>45%</td>
</tr>
<tr>
<td>Bicester Community College - Bicester</td>
<td>43%</td>
<td>48%</td>
<td>98%</td>
<td>63%</td>
<td>98%</td>
<td>39%</td>
<td>14%</td>
</tr>
<tr>
<td>Blessed George Napier Catholic</td>
<td>62%</td>
<td>62%</td>
<td>98%</td>
<td>78%</td>
<td>97%</td>
<td>59%</td>
<td>38%</td>
</tr>
<tr>
<td>School</td>
<td>Measure centred around 1000</td>
<td>Limit of Key Stage 2 to 4 CVA Confidence Intervals</td>
<td>Coverage</td>
<td>Number of qualifications</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----------------------------</td>
<td>----------------------------------------------------</td>
<td>----------</td>
<td>-------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Banbury School - Banbury</td>
<td>993.0</td>
<td>1001.8, 984.2</td>
<td>95%</td>
<td>10.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bartholomew School - Witney</td>
<td>987.4</td>
<td>998.0, 976.8</td>
<td>96%</td>
<td>11.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicester Community College - Bicester</td>
<td>994.3</td>
<td>1004.0, 984.7</td>
<td>97%</td>
<td>11.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blessed George Napier Catholic School and Sports College - Banbury</td>
<td>1001.0</td>
<td>1012.4, 989.5</td>
<td>94%</td>
<td>10.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Secondary schools in Oxfordshire

This is an alphabetical list. The columns show the proportion of pupils attaining five or more good GCSEs and equivalents at the end of Key Stage 4 including English and maths, the Level 2 contextual value added (CVA) score, the average A/AS-level and equivalent points and the Level 3 CVA.

Click the name of an institution for a page detailing its performance.

Click here for a fuller explanation of the figures.

<table>
<thead>
<tr>
<th>School</th>
<th>Sort by GCSE %</th>
<th>Level 2 CVA</th>
<th>Level 3 CVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banbury School</td>
<td>40</td>
<td>993</td>
<td>647.3</td>
</tr>
<tr>
<td>Bartholomew School</td>
<td>66</td>
<td>987.4</td>
<td>810.6</td>
</tr>
<tr>
<td>Bicester Community College</td>
<td>43</td>
<td>994.2</td>
<td>678.3</td>
</tr>
<tr>
<td>Blessed George Napier College</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catholic School and Sports College</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Limitation 1

Past performance is no guarantee of future performance ... in many cases the value of the investment can fall as well as rise
Seven years out of date!

- The 2009 school league table report schools’ performances for the 2009 GCSE cohort

- However, parents want to know schools’ performances for the 2016 GCSE cohort

- Inferences about the future performances of schools will be far less precise than inferences about their current performances
The CVA model is a two-level multilevel model

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \]

\[ u_j \sim N\left(0, \sigma_u^2\right), \quad e_{ij} \sim N\left(0, \sigma_e^2\right) \]

- \( y_{ij} \) is the total age 16 GCSE score for student \( i \) in school \( j \)
- \( x_{ij} \) is their average age 11 English and maths score
- \( u_j \) is the **CVA school effect** for secondary school \( j \)
- \( e_{ij} \) is the student level random effect or residual
Data

• National Pupil Database (NPD)

• We focus on the 2009 GCSE cohort

• We analyse a 10% random sample of schools
School effects for the 2009 cohort

- ~60% of schools are significantly different from the overall average
School effects for the 2016 cohort

• Will the same significant differences remain in 2016?

• We must factor in the additional uncertainty that arises from predicting seven years into the future

• We use a multivariate response version of the CVA model for eight cohorts of students to do this
Multivariate response model for all eight cohorts: 2002-2009

- These correlations measure the **stability of school effects** over time

<table>
<thead>
<tr>
<th></th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>0.90</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>0.82</td>
<td>0.90</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>0.75</td>
<td>0.82</td>
<td>0.90</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>0.69</td>
<td>0.75</td>
<td>0.82</td>
<td>0.90</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>0.62</td>
<td>0.69</td>
<td>0.75</td>
<td>0.82</td>
<td>0.90</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>0.58</td>
<td>0.62</td>
<td>0.69</td>
<td>0.75</td>
<td>0.82</td>
<td>0.90</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>0.55</td>
<td>0.58</td>
<td>0.62</td>
<td>0.69</td>
<td>0.75</td>
<td>0.82</td>
<td>0.90</td>
<td>1.00</td>
</tr>
</tbody>
</table>

- The **seven-cohort apart correlation** is just 0.55
School effects for the 2009 cohort vs. the 2002 cohort

- The correlation of 0.55 implies a substantial reordering of schools.
- The government implicitly assume that there is no reordering.
Comparison of the school effects for the 2009 and 2016 cohorts

School effects for the 2009 cohort

More appropriate for inferences about school accountability

Predicted school effects for the 2016 cohort

More appropriate for inferences about school choice

- Different users want different things from league tables
Making more precise predictions?

• We have used 2009 data to predict 2016 performance

• What about using data from 2008, 2007,...?
  – Note that earlier cohorts will add increasingly less information
Making more precise predictions?

- There is no visible improvement in the precision of the predictions
Limitation 2

Should we adjust for school level variables?
Adjusting for school level variables

- The government adjust for two school level variables
  - **School mean** of age 11 intake achievement
  - **School spread** of age 11 intake achievement
    - There is a positive effect of having a high achieving and homogenous intake
    - These are school composition variables that aim to measure peer group effects
- Including these variables removes peer group effects from schools’ measured performances
Different users want different things from league tables

• For choosing a school:
  – No adjustment should be made
  – Parents are interested in how much better their child will do in one school than another
  – Peer-group effects are part of the difference between schools which is of interest

• For holding schools accountable:
  – This adjustment should be made
  – Schools should not be held accountable for factors outside their control
  – The government is interested in disentangling schools’ policies and practices from their context and peer groups (this is ambitious!)
Adjusting and not adjusting for school compositional variables

Adjusting for the positive effects of having a high achieving and homogenous intake lowers the rankings of selective schools.

However, selective schools’ rankings will be lowered by too great an extent if selective schools are effective in their own right.

Being a selective school is confounded with having a high achieving and homogenous intake.
Other statistical limitations
Other statistical limitations

- At GCSE, students take **different combinations of subjects**
- Schools will be **differentially effective** for different types of students and for different responses
- **Student mobility** between schools is not recognised
- Students with **missing data** are listwise deleted
- Little is known about the **inter-rater reliability** of the tests
- For school choice, CIs for **multiple comparisons** are needed
Some broader limitations

- Huge financial cost to implement
- Teaching time is taken up with the administrative burden of the tests
- The range of knowledge and skills that tests assess is very narrow
- Stress caused by over-testing turns children off education
Conclusions
Conclusions

• School league tables ignore the uncertainty in using current performance as a guide to future performance
  – Adjusting for this uncertainty reduces the number of schools that can be separated to almost none

• For school choice, don’t adjust for school-level factors, since this is part of the effect that parents are interested in
  – Adjusting for school achievement composition pushes selective schools down the league tables
Conclusions (cont.)

• We have shown that CVA scores contain very little information for choosing schools
  – This is just one more argument against their publication
  – However, the government insist that they are here to stay
  – In which case, strong health warnings are required
  – They should never be the sole basis of high-stakes decisions

• There is still an accountability role for CVA scores
  – But should only be used sensitively by experts
  – Can be used as a monitoring and screening device
  – However, it is not clear how to adjust for school compositional variables that are correlated with school policies and practices
Conclusions (cont.)

• The issues we have discussed are also relevant for primary school, post-16 schools and university league tables
  
  – Small size of primary schools makes estimated school effects even more imprecise
  
  – Universities are even harder to compare than schools due to lack of common curriculum and tests

• They are also relevant to other countries which publish school league tables to inform choice
  
  – Australia, Canada and the US
References


• [http://www.bristol.ac.uk/cmpo/plug/](http://www.bristol.ac.uk/cmpo/plug/)