School league tables: what can they really tell us?

School league tables are eagerly scanned by parents hoping to find the best school for their child, and by teachers hoping to see their school rise in the rankings. But do they tell parents what they need to know? **Harvey Goldstein** and **George Leckie** argue that league tables are not fit for that purpose, and say it is time that their publication should cease.

Since rankings of schools' examination results were first published in Britain in the early 1990s they have become an established feature of the educational landscape. Since then they have been joined by key stage test score rankings at the end of primary school. Although limited now to England, they have become more sophisticated, most notably by acquiring so called "value-added" elements. When the results for schools are released by the government each autumn they are widely published as rankings in the national and local newspapers and are used by schools in their promotional literature. Those institutions seen to be high or low in the rankings are often singled out for praise or condemnation; the implicit assumption is that the position in the rankings can be treated as a measure of the quality of their educational provision. How justified is that assumption? How are these rankings or "league tables" promoted? And how useful are they in practice, both to an authority seeking out strong or weak schools and to a parent seeking the best school for their child? We shall not be overly technical but advise any reader interested in the technical detail to consult Goldstein and Spiegelhalter[1].

Accountability and choice

There are two principal justifications for league tables, one based on notions of institutional accountability and the other on their use for making choices between schools.

Holding schools publicly accountable for the performance of their students in public examinations and for their key stage test scores is argued to be fair and give an incentive to schools to improve their "standards". When first introduced, simple school averages—"raw scores"—formed the basis for rankings. In these, perhaps not surprisingly, schools that recruited from the brightest and most advantaged pupils tended to emerge at the top of the rankings. In 1995 the government accepted the research evidence and agreed to move to a "value-added" system. This takes account of the differing intake achievements of students entering the school. Explicit or implicit selection procedures, for example, would affect the value-added score. More recently so called "contextual value-added" systems have been used which, in addition to adjusting for individual student prior achievements, also attempt to adjust for such factors as the average prior achievement of a student's peers. Eligibility for free school meals and lack of spoken English at home come into play here. Thus, an inner-city school with a so-called "sink estate" catchment area would receive more contextual value-added points.

The current rankings produced by this last system are used by the official school inspection system (OFSTED) to inform their judgements, and also, in some places, as part of a local or internal accountability *screening* system to identify those "outlying" institutions that may require special attention[2]. The interesting aspect of such "internal accountability" systems is that they

Do league tables give any guide to how well a school is performing?

treat the rankings as just one source of evidence about *potential* problems to be further investigated rather than as definitive statements about the quality of education within the institutions. Clearly, then, education officials do not regard rankings on their own as strong enough evidence for a judgement.

This is reinforced by recognising 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, the Department for Schools, Children and Families (DCSF) website (`http://www.standards.dfes.gov.uk/`) now provides intervals for value-added estimates, although these are generally not prominent in media presentations or discussions.

In its booklet of guidance for parents choosing schools the DCSF states: "The Government publishes these [league tables] every year. They tell you how well pupils did in exams at every school—and you can compare one school's results to others in your area and nationally" (`http://www.dfes.gov.uk/sacode/`). For parents, access to these data is typically through the tables published in the media. Yet, when making a choice of school the same issues occur as when the tables are used for accountability purposes. There are, however, two important differences.

From the point of view of school choice it seems clear that we should not adjust for any school level factors—those taken account of in the contextual value-added rankings. 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 therefore not something to be adjusted for. Thus, the DCSF contextual value-added estimates are *not* appropriate for choice purposes. They do indeed give different rankings from the raw scores, with a correlation of just 0.76 between the two sets of ranks. In terms of school choice there is a further key issue, which is entirely ignored in the tables currently produced. To illustrate, consider the case of secondary schools and results of GCSE examinations taken by pupils at age 16. For a cohort of 11-year-olds entering schools in 2008, the relevant GCSE exam results will be for the year 2013. In other words for the purpose of choice, what is required are *predicted* school-effects some 6 years beyond those typically currently available. Given that the correlation between school-effects for cohorts of children taking such exams 6 years apart is only about 0.6, this considerably reduces the precision of any comparisons. In other words, exam performance now is a poor guide to performance in 6 years time. In the next section we shall illustrate this with recent data.

Finally, for both accountability and choice purposes it has long been recognised that "differential" effects exist. Thus, for example, the differences between schools are known to depend on the prior achievement value so that rankings will change depending on this prior achievement chosen for comparison. Again, this is ignored in the current tables, but should be borne in mind.

In the next section we look in more detail at some actual rankings to illustrate our points.

Examples: choosing schools

We have analysed data for 2007 GCSE scores using the National Pupil database (NPD; see `http://www.bris.ac.uk/Depts/CMPO/PLUG/whatisplug.htm`). This is 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 have matched the examination scores to the key stage 2 exams in 2002.

The caterpillar plot in Figure 1 shows the rankings of 54 secondary schools in one local authority in 2004. The top graph shows the mean exam scores—the raw scores—ranked from lowest to highest with 95% confidence intervals. The bottom graph is similar but is based on the value-added school estimates. We have also highlighted two schools: in the top graph one, indicated by the large blue triangle, is significantly below the overall mean and the other, indicated by the large green circle, is significantly above (they are also significantly different from each other). However, their value-added estimates in the lower graph place them both near the centre of the distribution with no evidence that either is significantly different from the mean or from each other.

By publishing both the raw scores and value-added ones it is left open to schools to choose the data that shows them in the best light in their promotional literature. Users such as parents should therefore take care to discover which set of data is being used. Users should also try to obtain appropriate confidence interval estimates, although the situation is somewhat problematic since confi-

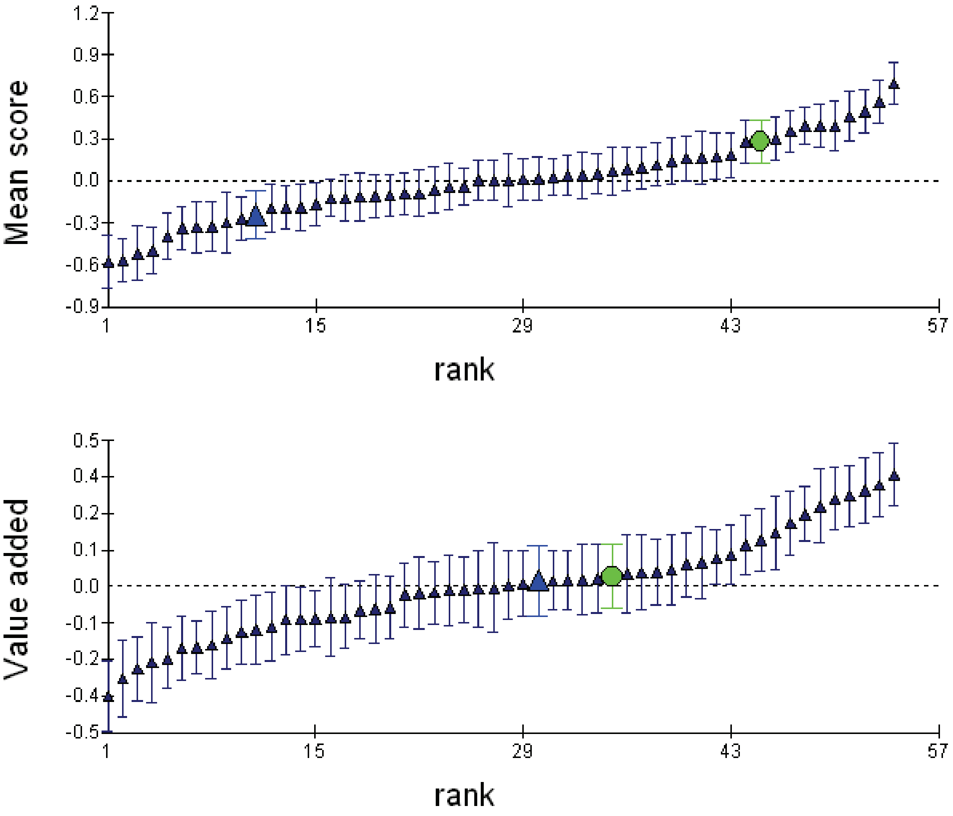

Figure 1. Rankings of 54 secondary schools from one local authority in 2004

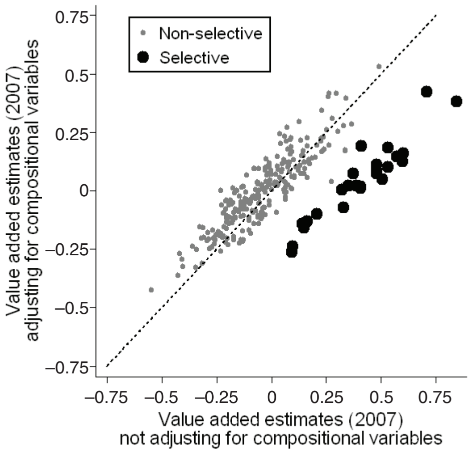

Figure 2. Contextual *versus* (non-contextual) value-added estimates

dence intervals for the raw scores are not made available by the DCSF.

In Figure 2 we analyse 274 secondary schools, which have been chosen at random from the 2007 data. We plot the rank of the contextual value-added scores for these schools against the rank of their value-added scores. In both cases we adjust for a range of pupil level variables including gender, age, free school meal status and ethnicity. For the contextual value-added model we additionally adjust for the mean and standard deviation of the prior test scores of all the pupils in each pupil's school. In the graph, schools have higher ranks the further they are from the origin and schools that lie above the 45° line perform better on the contextual value-added measure than on the value-added measure. The graph shows that adjusting for pupils' characteristics alters the rankings for many schools. Interestingly, it shows that for selective (grammar) schools the addition of compositional variables tends to lower the relative rankings. 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 are also adjusting for the initially higher prior achievements and smaller variability in this group of schools. This finding for grammar schools may be quite relevant for accountability purposes, when we wish to discount the initial advantages of such schools, but is not relevant for school choice.

Finally, Figure 3 is a value-added caterpillar plot using the data on the 54 schools as described above, but now displaying the predicted estimates 6 years ahead, in 2010, assuming a correlation of 0.6 between the estimates for 2004 and 2010 and the same variance between schools in each year. We see clearly that no schools can be separated from the average, nor are there any pairwise significant differences. We find a similar result for the larger data set of 274 schools where we estimate that under 5% of schools can be separated from the average using predictions 6 years ahead, compared to just over 60% that can be separated on the basis of current estimates alone.

Conclusions

We have argued that the publication of school league tables, as they currently stand, leaves much to be desired. It is generally recognised that the promotion of raw scores as judgements of school quality is not justified. In addition, we have argued that the use of value-added scores for school choice is severely constrained by wide confidence intervals and the extra uncertainty introduced when one considers that the person making a choice of school is actually interested in future predicted values. We have shown that taking this extra uncertainty into account implies that very few schools indeed can be separated from each other or, indeed, from the population mean. In addition, the present DCSF contextual value-added tables are inappropriate for school choice, despite being promoted as such. Parents relying on league tables to select a school for their children are using a tool not fit for that purpose.

In terms of accountability the inherent imprecision of all estimates reduces their usefulness for accountability other than for internal "screening" purposes or, when properly understood, to assist the judgements of school inspectors.

We have said nothing, since it was not our principal purpose, about the side effects and perverse incentives generated by the use of league tables. These are undoubtedly serious.

There is an incentive, for example, for a school to discourage pupils from taking "hard" subjects, such as foreign languages and sciences, because they fear depressing the proportion achieving passes. Some schools, it is said, concentrate excessively on "borderline" pupils, who might just scrape the C grade which counts towards the school's score, at the expense of those striving for A grades and those who might manage a D but not a C.

Now seems a good time to abandon school league tables.

The interested reader is referred to the report of a Royal Statistical Society working party for a detailed discussion[3]. Although we have not discussed the use of measures of trends over time—so called improvement scores produced by the DCSF—the same issues apply to these, especially since these scores are generally not based on any kind of value-added analysis.

Finally, it is noteworthy that Scotland, Wales and Northern Ireland have either never had or have moved away from publishing school league tables. Now seems a good time for England to follow suit.

References

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

2. 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.

3. Bird, S., Cox, D., Farewell, V. T., Goldstein, H., Holt, T. and Smith, P. C. (2005) Performance indicators: good, bad, and ugly. *Journal of the Royal Statistical Society, Series A*, **168**, 1–27.

Harvey Goldstein and George Leckie are both at the Centre for Multilevel Modelling, Graduate School of Education, University of Bristol.

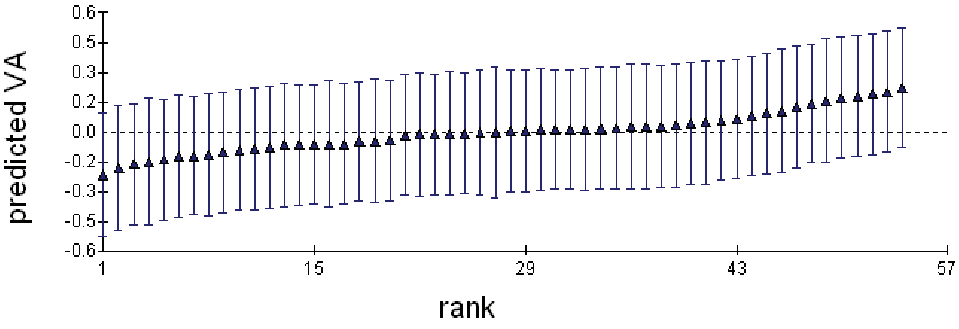

Figure 3. Ranks of value-added scores predicted 6 years ahead