Now you see it, now you don't: school effectiveness as conjuring?

Stephen Gorard, University of Birmingham

Introduction

I once watched an illusionist on television—Derren Brown—and he successfully predicted the winner of a series of horse races, so allowing a member of the public to collect on a series of wagers. The programme showed him ringing a woman up, explaining that he wanted her to place a bet on a specific horse running in a specific future race because his 'system' revealed that it would come first. She then made the bet and duly won, and the sequence was repeated. At first appearance, and assuming that the woman is not an accomplice and the whole thing is not staged, this seems to be an amazing trick or perhaps a staggering coincidence. Of course, it is neither. The illusionist has rung a series of people, ensuring that each horse running in any race is backed by someone at the expense of the programme, in an exercise involving thousands of people. The TV programme then only shows the inevitable winner. The illusionist has genuinely 'predicted' the winner of each race but this means nothing because of the number of attempts. The point of the story, for this brief paper, is that something that might look impressive in isolation may be less so once a fuller picture of the number of attempts is revealed.

Consider the purported identification of school effects via value-added analysis. The idea is to use existing data on all pupils in the relevant school population to predict as accurately as possible how well each pupil will score in a subsequent examination. The difference between the predicted and observed score in the examination is then used as a residual. The averaged residuals for each school are termed the 'school effects'—and are intended to represent the amount by which pupils in that school progress more or less in comparison to equivalent pupils in other schools. If a school has a systematically beneficial impact on a reasonable number of pupils the average residual will be positive, otherwise negative or zero. In England, the official calculation published by the Department for Education is termed contextualised value-added (CVA) because a large battery of context variables is used with measures of prior attainment to improve prediction of future examination scores. The official interpretation is that any school

CVA score markedly above or below zero represents an important and stable characteristic that can be used for policy and to improve practice. The results are published as 'School Performance Tables' (http://www.education.gov.uk/performancetables/, accessed 27/1/11), used by the school inspectorate OFSTED to help judge the quality of schools, used by schools themselves to target resources and improvements, and by some parents to help select a school for their child. The results matter.

The whole scheme is clever and well-intentioned. But are CVA results a fair test of school performance in the way that they are clearly designed to be? I have shown elsewhere that value-added fails in one major respect (Gorard, 2006a). The results are not independent of the absolute level of attainment, and so suffer the same flaw as the raw examination results they are intended to replace. And the fact that there is variation in the CVA scores between schools does not of itself make them meaningful. Even a dataset made up of completely random numbers can be fitted to a regression model post hoc with near 100% accuracy (Gorard, 2008). The variation in CVA scores between schools could easily emerge from the propagation of missing data and initial errors in the measurement of examination results and in the battery of contextual values used to improve the predicted score (Gorard, 2010a). Perhaps the results are not random, but they could be heavily biased so that they would still mean nothing even if systematic differences in outcomes recur. For example, the formula might over or undercompensate for free school meal pupil intakes, meaning that schools with high (or low) poverty levels among pupils always tend to be above average in CVA. The formula might err in not taking enough account of the proportion of girls, who tend to do slightly better than boys overall in any school, meaning that girls' schools might always tend to be above average in CVA. So, it is entirely possible that the variation between schools on this complex CVA score is meaningless, and that governments, schools and families are being misled in using it in real life.

CVA in Worcestershire

Probably the most common defence of the usefulness of CVA I have heard over the years is that it must mean something because a specific school or type of school has had a succession of positive scores. Perhaps because this claim is usually made by someone involved with the school or type of school in question, I have never heard the equivalent argument based on a run of negative scores! This argument mistakes consistency (reliability) for meaning (validity), but it is still interesting to consider further. I have asked the question before, many times; what would we expect CVA scores to look like if, as in the betting illusion at the start of the paper, they actually meant nothing at all (Gorard, 2010a)? This paper uses figures for one local education authority in England over five years to consider how impressive the contextualised value-added (CVA) scores for all secondary schools are, once a fuller picture is revealed.

The way the scores are calculated makes them zero sum, and 1,000 is added to the result, presumably to avoid having negative values. By definition and design, around half of all schools in England will have scores above 1,000 and half below, and the average of them all must be 1,000. In 2010, there were 29 state-funded secondary schools in Worcestershire local authority (selected as a case study simply because it is where I was at time of writing). Of these, 16 had a contextualised value-added (CVA) score for Key Stage 2 (age 11) to Key Stage 4 (age 16) progress that was above the national average, and 13 were below. However, as is common, many schools had scores that were only just above or below the average. The overall CVA score for Worcestershire, weighted for the number of pupils whose results were used in each school, was 1,000.78. Worcestershire is therefore almost exactly average in terms of school performance, as assessed by national CVA. The picture is very similar for every year in which CVA has been calculated by the Department for Education (DfE) in England. The average CVA for Worcester is around 1,000, with just over half of the schools having CVA scores just above 1,000 every year. How consistent are the scores for each school or, put another way, how common are those schools that can boast of consistent positive (or negative) CVA?

From 2006 to 2010, a total of 30 secondary schools are listed as being in Worcestershire, of which seven have one or more years in which no data is available (Table 1). These schools are not considered further, since it not possible to say whether their scores are consistently positive or negative over time (although we could say, even on the data available, that three are definitely not consistent).

Of the 23 remaining schools with five complete years of data, 12 had some years with published positive CVA scores and other years with negative scores (Table 2). These 12 schools are therefore deemed not to have consistent CVA scores, and are not considered further. Of course, it could be that Arrow Vale Community School truly got better over this period, and so the existence of both positive and negative CVA scores, almost in a trend, might be meaningful. But then this would mean we had to argue that Hanley Castle High School got better, worse and better again all in the five year

Table 1 Worcestershire schools with missing data

School name	*2006*	*2007*	*2008*	*2009*	*2010*
Baxter B & E College	986	—	975	996	989
Bewdley High School	1,001	—	992	988	982
Elgar Technology College	962	990	986	964	—
King Charles I School	1,000	—	982	976	1,001
Stourport High School	1,022	—	1,016	1,019	1,010
Tudor Grange Academy	—	—	—	—	1,001
Wolverley High School	1,013	—	988	974	997

Source www.education.gov.uk/performancetables

Table 2 Worcestershire schools with both positive and negative CVA

School name	*2006*	*2007*	*2008*	*2009*	*2010*
Arrow Vale Community	974	995	993	1,018	1,021
Christopher Whitehead	1,031	1,011	1,015	1,010	993
Droitwich Spa High School	1,005	1,012	1,003	991	972
Dyson Perrins CofE High	1,005	986	985	975	987
Evesham High School	1,024	1,014	1,018	1,017	992
Hagley Catholic High	993	993	1,007	1,003	1,009
Hanley Castle High School	999	1,011	1,003	994	1,001
Kingsley College	1,009	984	989	984	977
Malvern, The Chase	1,001	995	992	993	993
Tenbury High School	1,003	1,004	1,000	1,015	988
Trinity High School	1,003	994	991	1,001	1,021
Waseley Hills High	1,008	1,011	1,009	1,009	999

Source www.education.gov.uk/performancetables

Table 3 Worcestershire schools with apparently consistent positive or negative CVA

School name	*2006*	*2007*	*2008*	*2009*	*2010*
Bishop Perowne CofE	984	995	993	970	988
Blessed Edward Oldcorne	1,009	1,007	1,001	1,000	1,016
Haybridge High School	1,025	1,025	1,026	1,030	1,029
Martley, The Chantry High	1,020	1,014	1,042	1,030	1,015
North Bromsgrove High	995	991	971	983	994
Nunnery Wood High School	1,022	1,031	1,024	1,017	1,006
Pershore High School	1,008	1,002	1,009	1,005	1,004
Prince Henry's High School	1,025	1,015	1,006	1,013	1,015
South Bromsgrove	1,002	1,002	1,005	1,012	1,013
St Augustine's Catholic	1,034	1,033	1,023	1,019	1,009
Woodrush Community	1,003	1,009	1,012	1,022	1,002

Source www.education.gov.uk/performancetables

period. It seems easier, assuming that school effectiveness is not to be such a volatile phenomenon as to make it unusable in practice, to ignore these 12 schools.

The remaining 11 schools are slightly harder to assess. Given that the CVA results for Worcestershire are slightly above the national average, we would expect a reasonable number of these schools to have scores at or just above 1,000. This is what we find (Table 3). But how far away from 1,000 does a score have to be before it is of substantive importance? The Department for Education, on the incorrect advice of statisticians both professional and academic, presents the CVA scores for each year and school with 95% confidence intervals. These intervals are calculated and presented in error because there is no random element for them to be estimating, and the

strong assumptions for their use have not been met (Gorard, 2006b; Gorard, 2010b). DfE have been misled here, and the use of confidence intervals with population data is patently absurd (see Gorard, 2008). But if we take these confidence intervals at their face value they can give us an idea of how confident the DfE themselves are about whether the CVA scores are actually different from zero, by considering whether both extremes of the interval are either positive or negative. This is how DfE advise users of the performance tables, such as parents, to consider confidence intervals. For example, the reported CVA score for Woodrush Community School in 2006 was 1,002.9 (last row in Table 3), and the reported 95% confidence interval was from 993.6 to 1,012.2. In other words, 1,002.9 is so close to 1,000 (for the level of variation encountered in doing the calculation) that DfE cannot be sure whether the real score is positive or negative or even just average. The CVA for the same school in 2008 was 1,011.9, and the confidence interval was from 1,002.9 to 1,020.9. In 2008, therefore, DfE can be more confident than in 2006 that, whatever the real CVA score is, it is positive. Put another way, with the data and distribution of data observed, a figure of 2 or 3 above or below 1,000 is not enough for them to be confident that any school has a CVA that diverges from 1,000. But a figure of 12 or more away from 1,000 is treated and reported by DfE as being substantially different to 1,000.

Of the 11 schools in Table 3, nine were in the same position as Woodrush Community School in having years where the extremes of the confidence interval are either side of 1,000. Bishop Perowne could easily have had a positive CVA score in reality for 2007 and 2008, for example. Blessed Edward Oldcorne could easily have had a negative score in reality for 2008 and 2009. And so on. Only two schools had successive years of CVA that were substantially different from 1,000 in this sense. In all five years the reported scores for Haybridge High School and Sixth Form, and Martley, the Chantry High School were noticeably higher than 1,000. But there are still two obstacles before anyone starts congratulating them on having robust and meaningful positive CVA scores. The first, and more minor point, is that the coverage (percentage of relevant age pupils included in the calculation) varied from year to year. The lowest coverage for Martley was only 95% of pupils included in the calculation, and this was in 2008 which was also the year of highest CVA score. Most schools could improve their relative position in CVA by selecting and omitting the least flattering 5% of their pupil scores. This is not what was done here, but it may have happened inadvertently if the pupils who are hardest to trace, most likely to drop out, or who take the least traditional qualifications are also likely to be the least flattering for the school CVA score.

The second objection is more important and indeed represents the key message of this paper. Using a zerosum calculation means that half of the scores in England will be positive and half negative in any year, even if the scores do not mean anything. Again, assuming for the moment that the scores have no meaning, we would expect half of any large group of schools,

such as the half of all schools in England with positive CVA in the first year, to get positive scores the next year. This means that after two years we would expect around one quarter of all schools (half of half) to have successive positive scores, one quarter to have successive negative, and for the other half of schools to have had one year of positive and one of negative (in either order). After three years, we would expect 1/8 of all schools to have successive positive scores, after four years 1/16, and after five years 1/32. Similarly, after five years we would expect 1/32 of all schools to have had successive years of negative CVA scores. So, with 30 schools in Worcestershire, we would expect almost exactly two schools to have consistent positive (30 times 1/32) or negative (30 times 1/32) CVA from 2006 to 2010. If on the other hand, CVA is not meaningless but represents a relatively stable characteristic of school quality then we would expect to find many more than just two schools in this position. However unpalatable the message might be to Haybridge High School and Sixth Form, and Martley, the Chantry High School, their run of CVA scores is no more than we would expect in a 'race' run by 30 schools with a random result. Probably a more palatable message for Bishop Perowne CofE and North Bromsgrove High schools is that their run of apparently negative CVA scores might also mean nothing.

Conclusion

Only two schools in Worcestershire had clearly consistent CVA over five years, even using a weak interpretation of either positive or negative and ignoring the actual scale of the scores. So whatever CVA is scoring it is very volatile. It would be absurd to encourage parents to use purported 'school effects' to help select a secondary school for an 11 year old on the basis of its CVA for a Key Stage that will be five years away for a new entrant. As this paper illustrates, the school CVA five years later may be nothing like that at time of choosing. Now this is not proof that CVA is meaningless. But it should provide no comfort to those who defend CVA and similar measures as *meaningful*. After the event, just like after the horse race, it is easy to pick the winner. And before the event, just like before the race, it is easy to estimate how many CVA winners and losers there will be. There will be exactly $n/2^{y-1}$, where y is the number of successive years of CVA data and n is the number of schools. Finding two schools with robustly consistent positive or negative CVA scores over five years from 30 in Worcestershire is exactly how many we would predict.

The findings here are relevant beyond Worcestershire LEA, since there is no reason to assume that any other area with a reasonable number of schools will show anything different (perhaps some doctoral researchers would like to try further case studies). CVA is volatile, unreliable, and based on high levels of measurement error. This does not mean, of course, that schools do not make a difference to their pupils, or that they are not differentially effective. It just means that traditional school effectiveness approaches like

CVA, based on a central zero sum calculation, seem ineffective in picking this difference up. The purported school effects assessed by CVA could be as illusory as Derren Brown picking the winner of each horse race. They could be school effectiveness as a conjuring trick. And if this kind of 'school effect' is as transitory as this case study of one authority suggests, then it is not something that can form an ethical basis for policy or practice decisions. Parents should not be (or have been) encouraged to choose, nor inspectors to judge, schools on this basis, until the situation is much clearer.

References

Gorard, S. (2006a), 'Value-added is of little value', *Journal of Educational Policy*, 21 (2), 233–241.

—— (2006b), 'Towards a judgement based statistical analysis', *British Journal of Sociology of Education*, 27 (1), 67–80.

—— (2008), '*Quantitative research in education: Volumes 1 to 3*', London: Sage.

—— (2010a), 'Serious doubts about school effectiveness', *British Educational Research Journal*, 36 (5), 735–766.

—— (2010b), 'All evidence is equal: the flaw in statistical reasoning', *Oxford Review of Education*, 36 (1), 63–77.