International Education Studies; Vol. 6, No. 1; 2013
ISSN 1913-9020 E-ISSN 1913-9039
Published by Canadian Center of Science and Education

How Unstable are ‘School Effects’ Assessed by a Value-added Technique?

Stephen Gorard[1], Rita Hordosy[1] & Nadia Siddiqui[1]

[1] School of Education, University of Birmingham, B15 2TT, UK

Correspondence: Stephen Gorard, School of Education, University of Birmingham, B15 2TT, UK. Tel: 44-121-414-4828. E-mail: s.gorard@bham.ac.uk

Received: October 19, 2012 Accepted: October 25, 2012 Online Published: November 12, 2012

doi:10.5539/ies.v6n1p1 URL: http://dx.doi.org/10.5539/ies.v6n1p1

Abstract

This paper re-considers the widespread use of value-added approaches to estimate school ‘effects’, and shows the results to be very unstable over time. The paper uses as an example the contextualised value-added scores of all secondary schools in England. The study asks how many schools with at least 99% of their pupils included in the VA calculations, and with data for all years, had VA measures that were clearly positive for five years. The answer is - none. Whatever it is that VA is measuring, if it is measuring anything at all, it is not a consistent characteristic of schools. To find no schools with five successive years of positive VA means that parents could not use it as a way of judging how well their primary age children would do at age 16 in their future secondary school. Contextualised value-added (CVA) is used here for the calculations because there is good data covering five years that allows judgement of its consistency as a purported school characteristic. However, what is true of CVA is almost certainly true of VA approaches more generally, whether for schools, colleges, departments or individual teachers, in England and everywhere else. Until their problems have been resolved by further development to handle missing and erroneous data, value-added models should not be used in practice. Commentators, policy-makers, educators and families need to be warned. If value-added scores are as meaningless as they appear to be, there is a serious ethical issue wherever they have been or continue to be used to reward and punish schools or make policy decisions.

Keywords: value-added, school effectiveness, England, secondary schools

1. Introduction

1.1 Why Value-added is Used

Governments worldwide, education leaders, teachers and families would all like to be able to judge the performance of schools and teachers in terms of pupil attainment (Barber and Moursched 2007). They want to know how much schools and teachers contribute to pupil attainment, how well schools overcome differences between the socio-economic background of their intakes, and whether some schools are more effective than others with equivalent pupils. It is clear that the performance of schools and teachers cannot be accurately assessed in terms of the *raw-score* attainment of their pupils, since this may reflect merely the quality of the intake to the school. For example, grammar schools in England select those pupils at age 11 who are most likely to do well in formal qualifications at ages 14, 16 and beyond. If at age 16 the pupils in a grammar school get better qualifications than pupils in a nearby school that takes only those pupils not accepted for the grammar school, then this is evidence that the grammar school selected their pupils well. It is not evidence, of course, that the grammar school itself and its teachers performed better than in the other school. It is very possible that if the pupils had somehow been swapped between the schools at the start, while the schools and teachers remained the same, then the pupils now in the grammar school would have done worse. This is not to say that particular schools do not make a difference, but that a great deal of the difference between school outcomes is directly attributable to pupil intakes. And what is true about the pupil intakes to grammar and secondary-modern schools is also possible for all schools. Where a school is sited, its specialism, organisation and precise methods of allocating places to pupils mean that there is considerable variation in school intakes, in terms of prior pupil learning, and indicators of possible disadvantage (Gorard and Cheng 2011).

The value-added approach to judging school performance, which has grown in popularity and importance from the 1980s, was therefore a good idea (Rutter et al. 1979). Here, schools are judged by the progress that their pupils make during attendance at the school, not on their absolute levels of attainment (e.g. Lubienski and Lubienski 2006). Data on all pupils in the relevant school population is used to predict as accurately as possible how well each pupil will score in a subsequent test of attainment. Any difference between the predicted and observed test result is then used as a residual. The averaged residuals for each school are termed the school's "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. A school with an average residual of zero is estimated to be 'performing' about as well as can be expected, given its intake. A school with an average above zero is doing better than expected. And this judgement about progress is intended to be independent of the raw-score figures, making it fairer than assessment of raw scores. Since this 'school effect' is deemed a characteristic of the school, not its specific cohort of pupils, it should be reasonably consistent over time where the staff, structures, curriculum, leadership and resources of the school remain similar over time.

In England, and elsewhere, this approach took firm hold in policy and practice, and the results are in widespread use. National figures for all schools are published by the Department for Education as 'School Performance Tables'. Individual school results are used in setting targets, development plans, assisting the school inspectorate OFSTED to help judge the quality of schools, and by some parents to help select a school for their child (Evans 2008). They have even been used to close schools down. Disaggregated results have been used to reward (or caution) individual teachers or departments. The results therefore matter, and the assumption has generally been that the method works well enough to form the basis for such life-changing decisions.

1.2 Concerns over the Value-added Approach

However, there is growing evidence that the value-added method, for all of its appeal, does not work particularly well (Hoyle and Robinson 2003). Of course, on reflection, value-added (VA) scores could never be independent of raw-scores, since half of the variation in VA comes from the prior attainment scores and half from the final attainment scores. The R-squared correlation between raw-score attainment (both prior and subsequent) and the value-added scores for an individual pupil, teacher or school is and must always be at least 0.5. In practice, it is considerably higher even than this (Gorard 2006, 2008). So, VA may not be as fair as it sounds on first hearing, because it is heavily dependent on those raw-scores that have been rejected as a fair assessment of school effectiveness. The precise results are also heavily dependent on the quality and completeness of the data. In every VA calculation there can be pupils with final scores but no prior attainment recorded. For example, a child moving from a private school or from another country like Wales where there has been no Key Stage 2 testing will have no prior record in England. And there can be pupils moving in the opposite direction, with prior attainment but no final scores. In practice, the Key Stage 2 (primary) to Key Stage 4 (secondary) VA data has over 10% such un-matched pupil records every year (Gorard 2010). This creates an initial error component (source of inaccuracy and bias) of at least 10% in any VA calculation, and there is no way of adjusting for this statistically since the data just does not exist. It would be wrong to assume that the missing data was somehow a random subset of the data that does exist (Pugh and Mangan 2003, Amrein-Beardsley 2008).

The situation is worse than this in three important ways. First, even where the scores exist they cannot be assumed to be totally accurate (Lamprianou 2009). Creating valid, comparable and reliable attainment scores is fraught with difficulty (witness the annual round of appeals against grades and admitted mistakes in marking). Second, the system used in England from 2006 to 2010 factored pupil background characteristics into the calculations. This was done in order to improve the quality of the predictions and reduce the size of the residuals for disadvantaged groups of pupils (Evans 2008). Again, this Contextualised Value-added (CVA) sounds a sensible and fair innovation. But it means that more data is needed on each pupil and this adds considerably to the level of missing data. At least 10% of pupils are missing data every year on each key variable such as whether they are eligible for free school meals, living in care, their ethnicity or additional educational needs. Third, and perhaps most importantly, all of these initial errors are compounded by the VA calculation itself to generate a far higher level of error in the residuals that result. Put simply, the residual is the difference between the predicted and attained score for each pupil, and because the VA model is the best fitting one for each dataset, the residuals tend to be very small. They are small in comparison to the actual scores. But because the errors in the actual scores can be negative or positive, when the residuals are created their initial errors can be added. The outcome is a larger error component in a much smaller result. The maximum error can, and usually does, dwarf the residual by several orders of magnitude (Gorard 2010). This makes the estimated result just about meaningless.

All of this means that VA results cannot be taken on trust or theory alone. Given this, it is remarkable how long this school effectiveness approach has survived, and how seriously it has been taken by governments. More importantly there is nothing to calibrate VA scores with. Real measurements, such as lengths, with an agreed standard scale can be compared with that standard. This is a crucial step in calibrating instruments, for example. A ruler can be judged accurate or not against a standard length. There is no external referent of this kind for school effects. They are operationally defined, quite simply, as what a measure of school effectiveness (like VA) estimates. Without any ability to calibrate the 'results', school effectiveness estimates may be wildly inaccurate. They may not be a measure of anything at all. In our experience, practitioners and school leaders have tried to defend the meaningfulness of VA scores by pointing to the runs of positive scores in their own schools or areas (Gorard 2011). This is to confuse reliability with validity. Nevertheless, to be useful VA has to has be a relatively stable characteristic of a school. Yet prior research agrees that VA scores are not stable across years. For example, McCaffrey et al. (2009) found annual correlations for teacher VA scores of between 0.2 and 0.7. This means that less than half of the variance (and usually less than a quarter) is not common to successive years.

This paper looks seriously at the idea that the 'school effects' resulting from VA calculations are not genuinely a characteristic of schools themselves, by developing the case study of one local authority in Gorard (2011) into a national picture. 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) value-added scores?

2. Method

We downloaded the results for all schools classified as 'Secondary (GCSE and equivalent)' on the Department of Education, for England, Performance Tables website at http://www.education.gov.uk/performancetables/ (last accessed 27/1/12), other than those designated 'Special schools'. We downloaded the number of pupils per school at the end of their Key Stage 4 (KS4), and the figures for Key Stage 2 to KS4 contextualised value-added. The CVA figures included the CVA estimated measure for each school, the 95% confidence interval for each school (upper and lower CVA estimate), and the coverage of these figures (the proportion of the total KS4 pupils included in the CVA measure). There were a total of 4,015 secondary school entries. The way the scores are calculated makes them zero-sum, and 1000 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 1000 (positive CVA), and half below (negative). For fuller details see DCSF (2007). This data is made freely publicly available and so provides no immediate ethical or consent issues. The data is for the population of maintained secondary schools. There is no sampling involved, and so no place for sampling theory statistics or significance tests in what follows.

In addition to special schools, we further excluded all schools for which there was no CVA information at all – usually 'other independent schools', the classification for fee-paying private schools. We then placed in a separate file all of the remaining schools for which there was complete CVA information for all of the years in which CVA was calculated and published, from 2006 to 2010. The most recent 2011 Performance Tables have a different value-added measure which is not directly comparable to previous years. The file of schools was sorted into those that had CVA measures of 1,000 (the average) and above, and those with CVA below 1,000, for 2006. The two sub-groups where then sorted into those with CVA at/above or below 1,000 for 2007. This created four sub-groups. One group was all schools with two successive years of positive (1,000+) CVA, another was all schools with two successive years of negative CVA. The other two groups had positive CVA followed by negative, or vice versa. This process continued for all years, until there 32 sub-groups. Of these 32 groups there was still one with all years of positive CVA and another with all years of negative.

Scatterplots were drawn for school CVA estimated measures against both the number of pupils involved and the coverage of the proportion of pupils involved, in each year. Similarly, Pearson R correlation coefficients were calculated for CVA measures and both the number of pupils and the coverage of each school, in each year. Pearson R correlation coefficients were also calculated for the CVA estimates across years. Because of what these scatterplots and correlations revealed, the official 95% confidence intervals were used to identify schools with clearly positive or clearly negative CVA in each year. Clearly positive meant that the lower limit for the confidence interval was 1,000 or more. Clearly negative meant that the higher limit was less than 1,000. This meant that very small schools with CVA not far from, and so indistinguishable from, 1,000 were omitted from the ensuing calculations. Schools with low coverage, that omitted a high proportion of pupils from their CVA estimates, were also removed from the calculations. The results were calculated for schools with at least 99% coverage.

3. Results

There were a total of 4,015 secondary school or college entries on the DfE School Performance Website. Of these, 1,118 (28%) had significant amounts of relevant information missing. Some had no information relevant to their performance at all. There are a number of good reasons for this, including schools that opened or closed between 2006 and 2010. The 2,897 schools that were initially retained for analysis had a valid entry for their CVA measure for all five years, even if they did not have full information on the number of pupils, the CVA confidence intervals, or the completeness (coverage) of the CVA scores. These schools are used as the basis for the ensuing calculations.

The first result is expected given previous studies in this area, and it is that individual school CVA scores are volatile over time (Table 1). Four years after the earliest score its common variance with the most up-to-date score is only around 21% (R=0.46). Even in one year, the common variance with the previous year is as low as under 35% (R=0.58). As will be shown below, this low level of stability over a very short period could be wholly explained by stability in the number of pupils (and so the confidence intervals used by the Department for Education) or in the percentage of pupils included in the calculation by any school (the coverage). Once these are adjusted for, there is almost no common variance between the CVA scores of five successive years.

Table 1. Correlation coefficients for CVA scores over time, England

	2006	2007	2008	2009	2010
2006	1	0.79	0.51	0.56	0.46
2007		1	0.58	0.67	0.56
2008			1	0.59	0.48
2009				1	0.79
2010					1

N= 2,897

If, for the sake of illustration, the CVA measures of these 2,897 schools meant nothing and are imagined to be a fluke or bias arising from errors in the data, then how many schools would have positive (or negative) CVA? In each year, almost exactly half of all schools will have positive (and negative) CVA because that is how CVA is calculated. If the CVA measure was really a fluke or otherwise not a consistent measure of the school to which it is attached, then in each succeeding year almost exactly half of those with positive scores the previous year will have a positive score again. After two years, around 25% of schools will have had two positive scores, 25% will have had one positive followed by one negative, and so on. Thus, after five years 2006 to 2010, 1/32 schools would be expected to have consistently positive scores each year. This would be 91 out of 2,897. And the same number would be expected to have five successive negative CVA scores, yielding a national total of 182 schools with a consistent direction of CVA, even if CVA was not actually a measurement of any school characteristic at all. Therefore, in order to establish that CVA means anything of consequence the number of schools with consistently positive or negative CVA after five years must be markedly greater than 182.

Before simply reading off the number of schools with consistent CVA from 2006 to 2010, there are a number of other data issues to deal with. Figure 1 shows the CVA scores for all schools (y axis) clustered around an average of 1,000, and the number of pupils in the calculation for each school (x axis). It is quite clear that the schools with the most extreme CVA scores, especially the positive ones, are among the smallest in England. The largest schools are much more tightly clustered around the mean CVA measure (at or near 1,000). This suggests that at least some of the CVA results are a consequence of the volatility of small numbers. The correlation between school size and the absolute deviation in CVA from 1,000 in 2010, for example, is -0.22. The highest measures of CVA, especially, tend to occur in small schools. But the same phenomenon occurs with the lowest CVA as well, as Figure 1 demonstrates. The correlation is not large, but because it applies in both directions of difference, it does suggest that at least some of the CVA measure is illusory, stemming from the volatility of small numbers. It also suggests that some of the apparent stability of CVA over years could be due to the stability of school size over successive years.

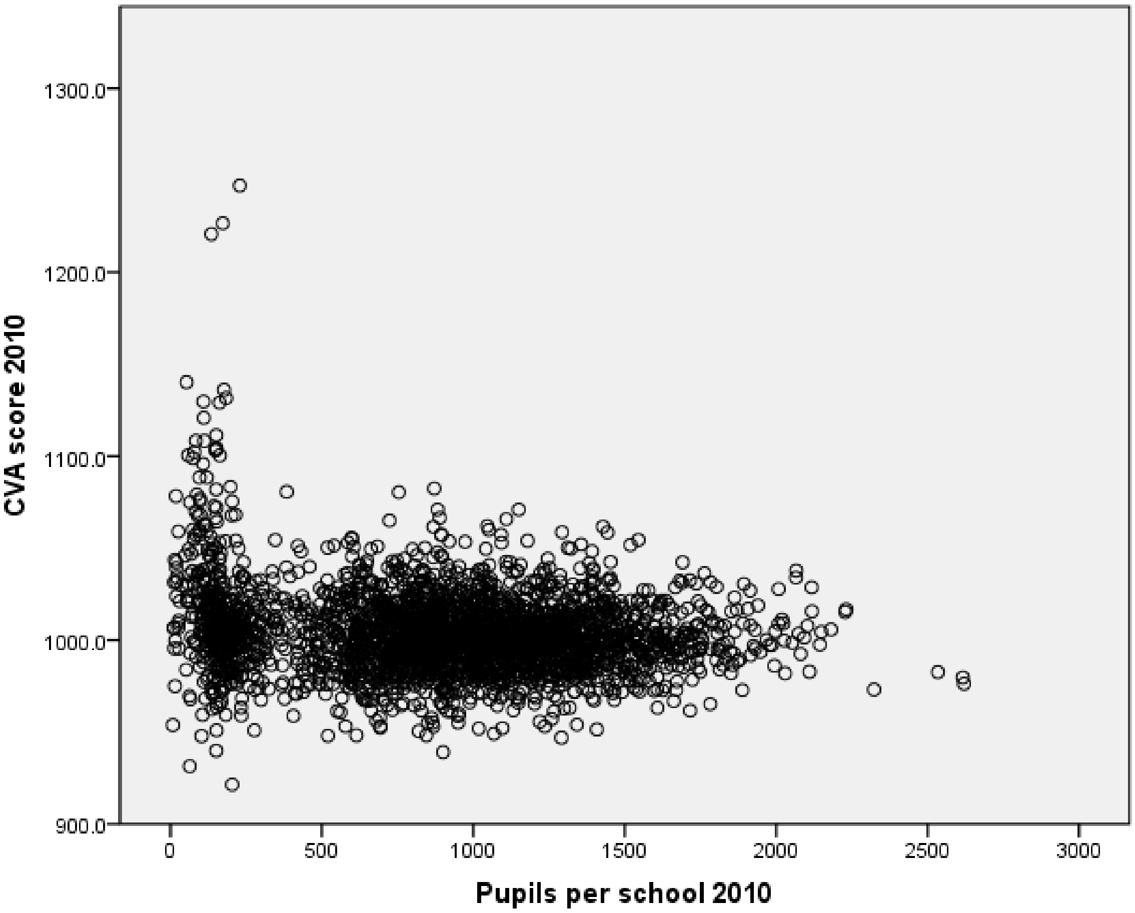

Figure 1. Crossplot of CVA measures and the number of pupils in each school, England, 2010

Figure 2 shows the CVA scores for all schools (y axis) clustered around an average of 1,000, and the proportion of the pupils in each school used for the CVA calculation (x axis). This should be 100% (1.00) but many schools have some pupils not included in their CVA calculations. This means that there is a slight tendency for schools with less than 100% of the data for their KS4 pupils to appear to have CVA scores diverging more from 1,000. For example, the school with the highest apparent CVA score in England has nearly 20% of its pupils missing from the calculation. It would, presumably, be possible for most schools to obtain an apparent positive CVA score if they could somehow eliminate from consideration the 20% least favourable pupils. The correlation between coverage and CVA distance from 1,000 was -0.21 in 2010. To some extent then, low coverage is the same as having fewer pupils at KS4. A CVA measure calculated for a year group of 10 with 100% coverage is based on the same number of pupils as one calculated for 20 pupils with 50% coverage. Having fewer pupils in a school calculation is linked to volatility, especially towards higher CVA (Figure 2).

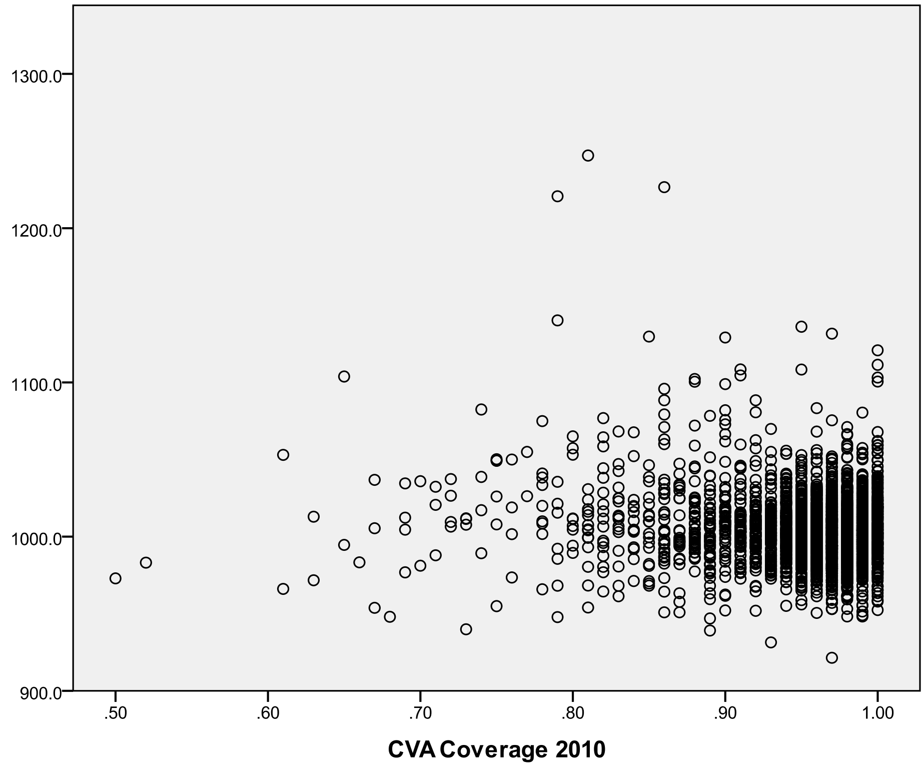

Figure 2. Crossplot of CVA measures and proportion of pupil coverage, England, 2010

The extent to which pupils are included in CVA calculations for each school (‘coverage’) is volatile over time (Table 2). The proportion of pupils whose scores are included in the CVA measure for each school is so variable that it is completely unrelated between years 2006 and 2007. It is also unrelated from 2006 and 2007 to any of the years 2008 to 2010. In 2009 and 2010 around half of the variation in coverage between schools is predictable from coverage the previous year. Put another way, until 2008 coverage by schools was completely erratic, while from 2008 onwards some schools were clearly either good or bad at providing data on all of their pupils. Coverage in both 2006 and 2010 was as low as 50% in some schools. This means that the problem is not the same as just having low numbers, and so volatility, in the KS4 year. If schools only have data on 50% of their pupils, there is a real danger of considerable selection bias. Most schools could improve their relative position in CVA enormously by selecting and omitting the least flattering 50% of their pupil scores. This 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. Even more significantly, this potential for bias is not addressed by the published confidence intervals for CVA. The confidence intervals only address the number of pupils and the size of the divergence of any school’s CVA from 1,000. They take no account at all of coverage problems.

Table 2. Correlation coefficients for Coverage over time, England

	2006	2007	2008	2009	2010
2006	1	0	0.01	-0.1	-0.01
2007		1	0.01	-0.01	0.01
2008			1	0.67	0.66
2009				1	0.73
2010					1

N= 2,897

The first step in dealing with these issues in the data is to reduce the apparent variation caused by the volatility of small numbers. One approach would be to eliminate from consideration as candidates for stability all schools below a certain size (100 pupils perhaps), but this ignores the scale of the CVA measurement. The judgement on inclusion needs to balance the scale of a school's deviation from 1,000 in CVA with its size. The DfE present 95% confidence bounds to handle this. The larger the school and/or the larger the deviation in CVA from 1,000, the more likely it is that the CVA measure and both the upper and lower confidence intervals will deviate from 1,000 in the same direction. The DfE are incorrect in using confidence intervals for this purpose, not least because there has been no random sampling of schools and so no sampling variation to estimate (Gorard 2011). Nevertheless, the confidence intervals do take both factors of scale into account. They are presented by DfE, and are used by schools and authorities as a way of judging whether the CVA deviation from 1,000 is 'significant' rather than a result of uncertainty and small numbers. Therefore, the rest of this paper considers only those schools whose entire 95% confidence interval is either above or below 1,000 as being positive or negative. Schools with CVA measures at or very near 1,000, and some very small schools, will be considered neither positive or negative. Since most school CVA measures are at or very near 1,000, because the CVA model was fitted by the DfE in England to be best fitting *post hoc*, there are only 443 schools remaining as candidates for stable positive or negative scores.

To deal with the coverage problem, and gain an estimate of its likely impact on CVA estimates, the analysis omitted all schools with less than 99% coverage. This is still a fairly tolerant approach, because almost any school could improve its CVA by omitting their least favourable 1% of pupils. However, there are only *three* schools in all of England with this level of data quality (Table 3). These schools are to be congratulated not denigrated. Every other school has at least one year with lower than 99% coverage. One of the three remaining schools had years of both positive and negative CVA. Both of the others had consistently negative CVA. There were *no* schools at all with consistently and clearly positive CVA that also included at least 99% of their pupils in the calculation. This is one approximate answer to the question about how many school from 2,897 have a consistent direction of CVA after five years. If CVA were meaningless, the answer would be 182 in theory (see above). In practice, the answer is 2. Since the use of confidence intervals eliminates as candidates those very small but high-CVA schools, and because 99% is still not complete coverage, and because we cannot include in the calculation any school that has no data in any years, this answer of two suggests that CVA does indeed mean nothing and has no practical value.

Table 3. CVA scores over time, schools with 99% coverage, England

	2006	2007	2008	2009	2010
Chasetown Sports College, Staffordshire	976	986	989	1016	1020
The Coleshill School, Warwickshire	980	974	978	962	973
Cheadle High School, Staffordshire	976	981	970	986	966

If we relax the coverage to a requirement of only 95% coverage in each year, there are still only 203 schools that had complete CVA measures for five years, with CVA in any year that was clearly above or below 1,000, and 95% coverage or better in all years. Of these 203, there were exactly 100 schools with CVA above 1,000 for five years, and 73 below. This 173 is another approximate answer to the question about how many school from 2,897 have a consistent direction of CVA after five years. It is still less than expected even if CVA were meaningless.

4. Conclusion

As Leckie and Goldstein (2009) and others have already noted, within a year or so the majority of a school's value-added score is unrelated to its prior VA. What this paper shows is that after five years, there is no clear relationship between the initial and eventual VA scores of a large number of schools. How volatile do VA scores have to be before we accept that they are meaningless with current datasets as well as useless or worse than useless for practical purposes? In any year, most schools have CVA scores indistinguishable from zero, but the same schools have other years where the CVA scores are apparently positive or negative. All of these cases have unstable CVA. Of the rest - that minority of schools with an apparently positive or negative CVA in any year - most of these same schools will have other years with CVA scores indistinguishable from zero (or even apparently opposite in sign). All of these cases also have unstable CVA. Of the remaining fraction of schools, many are small and/or with low pupil coverage in their CVA scores. Some schools have 50% or more of their pupils missing from their CVA calculations, for example. This makes any claim to stability for these cases scientifically unsound. Out of 2,897 schools used in the analysis here, only two had 99% or more pupil coverage for all years, with five successive years of apparently negative CVA. No schools had five years of apparently positive CVA.

This makes CVA almost entirely useless for practical purposes because it is not a consistent characteristic of schools. CVA could be something that changes in schools very rapidly, so that parents cannot rely on using it, when their child is choosing a secondary school at age 10 or 11, to make predictions about the likely results five or six years hence. Alternatively, the CVA scores could be largely not a characteristic of schools at all. It could be a measure of data quality or lack of data quality. Yet its use as a supposedly valid measure of school effects continues, even in England. CVA, the example used in this paper, is not the problem. CVA was a well-meaning and clever attempt to identify school effects. It is used here because of the quality of the publicly available data over five year for all maintained schools in England. Yet even with this high quality data VA does not seem to work. The problem lies with VA approaches more generally, including and perhaps especially for considering teacher effectiveness. The annual volatility of scores has been noted before, but not its extent after five years. And the implications not just for practice but for the field of school effectiveness have been largely ignored by policy-makers world-wide.

It is important to recall that the evidence in this paper does not suggest schools and teachers make no difference. Going to school is an important formative experience for young people, for good and ill. Going to school is very different to not going to school. However, the school system in England is designed through its funding, its laws about when and how school places are allocated, regulations about teacher development, inspections, national curriculum, and standard attainment in key stages, to try and make as little difference between schools as possible. And this is rightly so. The quality of education available in a national school system should not depend upon where a pupil lives. Perhaps what this study shows is that in this respect the system is working well. It does not matter much which school pupils attend. The evidence in this paper does not even mean that some schools are not more effective than others, with equivalent pupils. They may be. The paper merely suggests that traditional school effectiveness approaches based on VA, like the CVA model used in England, involving a central zero-sum calculation, are currently ineffective in picking this difference up. What VA models are mostly picking up instead is at least partly due to variation in the raw scores (Gorard 2006), partly due to a very large eventual error component (Gorard 2010), and partly due to factors like coverage and small numbers as reported here for the first time. Of course, readers must realise that the failure of VA approaches to solve the inherent unfairness of working with the more stable raw-scores of school-level attainment does not mean in any way that raw-scores should be used instead.

It is also important to recall that the assumptions made in this paper are not particularly strict. In assessing the consistency of VA scores over time, we have looked only at their direction, and whether a purportedly good school remains a ‘good’ school the following year. The scale of VA scores from year to year will be even more variable. As far as we can see there is no school that has exactly the same CVA score even across two years. Bizarrely, the precise model used for CVA calculations by the DfE each year was always modified by them once the results were in, to keep it as best fit. This means, for example, that the regression coefficient used to adjust for pupil poverty varied from year to year. This is strange. If prior attainment and pupil background characteristic make a difference to subsequent attainment, as they appear to, why is this difference not the same every year? The number of centimetres per metre, the boiling point of water, and Newton’s laws of mechanics are not adjusted every year or every use. Perhaps part of the inconsistency of school CVA scores over time lies in the inconsistency of the model created each time. The paper uses only 95% confidence intervals, as promoted by DfE, and insists only on 95% pupil coverage in the data. Yet, even with this approach of looking only at direction, with uncertainty in CVA measures, and up to 5% missing pupils, there are fewer schools with consistent CVA over five years than would be expected by chance alone.

As this paper illustrates, the school CVA five years later will most likely be nothing like that at time of choosing. Happily, CVA has ceased in England, and the 2011 and onwards School Performance Tables do not contain it. However, the tables do still have an entry for a simpler value-added score, based on a more prescriptive set of curriculum subjects. It is too early to test this for consistency, but all VA is inherently unstable (Mangan et al. 2005, Kelly and Monczunski 2007). It is therefore wrong for DfE to use it in public until it has been tested. The results in this paper mean that CVA, and by implication all VA models, cannot currently be used as an ethical basis for policy or practice decisions. Whatever it is that VA is scoring, with the quality of data available, it is so volatile that it would be absurd to encourage parents to use purported ‘school effects’ to help select a secondary school for an 11 year old (Leckie and Goldstein 2009). All of those schools and individual teachers negatively affected by their VA results may have been treated unfairly. Parents should not have been encouraged to choose, nor inspectors to judge, schools on this basis.

Acknowledgements

The authors are grateful to the British Academy for funding to support this work.

References

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system, *Educational Researcher*, *37*(2), 65-75. http://dx.doi.org/10.3102/0013189X08316420

Barber, M., & Mourshed, M. (2007). *How the world's best-performing school systems come out on top*, McKinsey & Co.

Coffield, F. (2012). Why the McKinsey reports will not improve school systems, *Journal of Education Policy*, *27*(1), 131-149. http://dx.doi.org/10.1080/02680939.2011.623243

DCSF. (2007). *A technical guide to the contextual value added 2007 model*. Retrieved from http://www.dcsf.gov.uk/performancetables/primary_07/2007GuidetoCVA.pdf, accessed 16/12/08

Evans, H. (2008). *Value-added in English schools*, London: Department for Children, Schools and Families. Retrieved December 16, 2008, from http://www.wcer.wisc.edu/news/events/VAM%20Conference%20Final%20Papers/VAMinEnglishSchools_HEvens.pdf

Gorard, S. (2006). 'Value-added is of little value'. *Journal of Educational Policy*, *21*(2), 233-241. Retrieved from http://taylorandfrancis.metapress.com/link.asp?target=contribution&id=J28232WJ836R5161

Gorard, S. (2008). The value-added of primary schools: what is it really measuring?, *Educational Review, 60*(2), 179-185. http://dx.doi.org/10.1080/00131910801934185

Gorard, S. (2010). 'Serious doubts about school effectiveness', *British Educational Research Journal*, *36*(5), 735-766. http://dx.doi.org/10.1080/01411920903144251

Gorard, S. (2011). Now you see it, now you don't: School effectiveness as conjuring?, *Research in Education*, *86*, 39-45

Gorard, S., & Cheng SC (2011). Pupil clustering in English secondary schools: One pattern or several?, *International Journal of Research and Method in Education*, *34*(3), 327-339. http://dx.doi.org/10.1080/1743727X.2011.609548

Hoyle, R., & Robinson, J. (2003). League tables and school effectiveness: a mathematical model, *Proceedings of the Royal Society of London B*, *270*, 113-199. http://dx.doi.org/10.1098/rspb.2002.2223

Kelly, S., & Monczunski, L. (2007). Overcoming the volatility in school-level gain scores: a new approach to identifying value-added with cross-sectional data, *Educational Researcher*, *36*(5), 279-287. http://dx.doi.org/10.3102/0013189X07306557

Lamprianou, I. (2009). Comparability of examination standards between subjects: an international perspective, *Oxford Review of Education*, *35*(2), 205-226. http://dx.doi.org/10.1080/03054980802649360

Leckie, G., & Goldstein, H. (2009). *The limitations of using school league tables to inform school choice*, Working Paper 09/208, Bristol: Centre for Market and Public Organisation

Lubienski, S., & Lubienski, C. (2006). School sector and academic achievement: a multi-level analysis of NAEP Mathematics data, *American Educational Research Journal*, *43*(4), 651-698. http://dx.doi.org/10.3102/00028312043004651

Mangan, J., Pugh, G., & Gray, J. (2005). Changes in examination performance in English secondary schools over the course of a decade, *School Effectiveness and School Improvement*, *16*(1), 29-50. http://dx.doi.org/10.1080/09243450500114843

McCaffrey, D., Sass, T., Lockwood, J., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates, *Education Finance and Policy*, *4*(4), 572-606. http://dx.doi.org/10.1162/edfp.2009.4.4.572

Pugh, G., & Mangan, J. (2003). What's in a trend? A comment on Gray, Goldstein and Thomas (2001), *British Educational Research Journal*, *29*(1), 77-82. http://dx.doi.org/10.1080/0141192032000057384

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books.