Educ Asse Eval Acc (2010) 22:181–198
DOI 10.1007/s11092-010-9100-4

Value-added measures for schools in England: looking inside the 'black box' of complex metrics

Anthony Kelly · Christopher Downey

Received: 8 July 2009 / Accepted: 19 May 2010 /
Published online: 28 May 2010

Abstract Value-added measures can be used to allocate funding to schools, to identify those institutions in need of special attention and to underpin government guidance on targets. In England, there has been a tendency to include in these measures an ever-greater number of contextualising variables and to develop ever-more complex models that encourage (or 'impose') in schools a single uniform method of analysing data, but whose intricacies are not fully understood by practitioners. The competing claims of robustness, usability and accessibility remain unresolved because it is unclear whether the purpose of the measurement is teacher accountability, pupil predictability or school improvement. This paper discusses the provenance and shortcomings of value-added measurement in England (and the Pupil Level Annual Schools Census that informs it) including the fact that although the metrics are essential for School Effectiveness Research, they fail to capture in its entirety the differential effectiveness of schools across the prior attainment range and across sub-groups of students and subjects.

Keywords Value-added measures · Pupil attainment data · Accountability in education · Complex models in school effectiveness

1 Introduction

Pupil-level data emerging from the introduction of the National Curriculum in England in 1988 enabled models to be developed of the value added (VA) by schools to the educative process. Until then, pupil attainment data came only from GCSE[1] examinations. The VA model currently in most widespread use in England—a *contextualised* value-added (CVA) model that takes account of school intake and

[1] Typically taken by 16-year old pupils in England after five years of secondary schooling.

A. Kelly (✉) · C. Downey
School of Education, University of Southampton, Highfield, Southampton SO17 1BJ, UK
e-mail: a.kelly@soton.ac.uk

situation—is one provided by the government (Department for Children, Schools and Families, DCSF), though increasingly since 2001, schools have additionally been using models developed by the independent Fischer Family Trust (FFT).[2]

While we (and most school effectiveness researchers) acknowledge the crucial benefit that VA measures bring in capturing the progress made by students in schools, we remain critical of the fact that, at practitioner level, the current model is caught in the tension between having to provide *both* key public accountability measures *and* data used to inform and steer improvement. We feel this tension raises 'practice concerns' around three key areas: the choice of contextualising factors included; issues around data and statistical modelling; and policy. Our concerns stem from our experience working with a wide range of practitioners engaged in school improvement in local authorities and schools, and in particular from a two-year funded research project on the use of data for school improvement conducted in more than 150 primary and secondary schools in England.[3]

After providing a brief history of the development of VA measures in England, we give a detailed treatment of our three areas of concern: the choice of contextualising factors; issues around data and statistical modelling; and policy. In each section we detail those aspects of the model that pertain directly to the issue under consideration and their implications, identifying in each case the tensions inherent in using the same CVA performance metrics for both public accountability and school improvement purposes.

2 Background: The development of value-added measures in England

Public reporting of secondary school performance was first introduced in 1992, alongside the creation of the government's Office for Standards in Education (Ofsted 2007). VA measures were based originally on seminal work in the field of school effectiveness by (amongst others) Fitz-Gibbon (1985), Aitkin and Longford (1986), Gray et al. (1986) and Willms (1992), but since then, the tables have become more complex as assessment results from the different Key Stages[4] (KS) have come on stream. In 2002, the first VA measures for secondary schools were included, based on school performance and assessment data ('PandA'), which Ofsted then used to drive its accountability agenda. As data from the various Key Stages was collated, it became possible to link prior attainment to outcomes and in 1998 the 'Autumn Package' was introduced to use this information for improvement purposes.

The underpinning rationale for value-added measures lies in the fact that without them, the achievement of schools could not take account of intake. Measures are

[2] The statistical methodology adopted by both is similar, although in the FFT multilevel model, students are allocated to one of five expected outcome bands based on prior attainment. FFT measures are calculated at three levels of contextualisation and utilise a slightly different set of factors from those employed in the DCSF model. Details are available from the 'Training Resources' section of the Trust's website (www.fischertrust.org).

[3] Joint ESRC/DTI-funded KTP, programme no. 1076.

[4] Compulsory schooling in England is divided into 'Key Stages': KS1 (Years 1 & 2) for ages 5–7; KS2 (Years 3–6) for ages 7–11; KS3 (Years 7–9) for ages 11–14; KS4 (Years 10 & 11) for ages 14–16. At each Key Stage, all children in state schools will study certain subjects, following the requirements of the National Curriculum.

both longitudinal (from Key Stage to Key Stage) and comparative (in the sense that *individual* pupil progression rates are compared with *national* progression rates). To succeed as both in a practitioner (as opposed to an academic) setting they must be fair and consistent over time and between schools, and to fulfil their school improvement function they must be well understood by those charged with planning and setting targets.

One major advantage of value-added measures over *threshold* performance indicators (such as the percentage of pupils attaining five or more A*–C grades at GCSE) is that *all* students contribute, rather than just those who happen to have crossed an arbitrary threshold, which lessens the temptation for schools to focus on borderline students at the expense of those who have predicted outcomes well below or safely above the threshold; and because they adjust for prior attainment, they are an improvement on raw measures of attainment such as 'Average Points Scores' (APS), which also factor in the contribution of every student. This notwithstanding, critics of early VA measures, such as Gorard (2006), questioned the value of the new measures. Having examined data from the DfES 2004 national School Performance Tables—in particular the correlation of KS2-4 VA scores for English secondary schools with the percentage of students from the same schools achieving five or more GCSE passes at grades A*–C (the then benchmark of raw attainment used to judge the performance of pupils and schools)—Gorard suggested that at least 70% of the variation in VA scores could be explained by the percentage of 5+A*–C passes at GCSE. His conclusion was that value added was 'of little value'.

Even as the first VA measures were being published, exploratory work was being carried out by statisticians working at the DfES (as the DCSF was then called) to incorporate a greater degree of contextualisation by adjusting for a number of additional pupil- and school-level demographic factors. This was made possible by the introduction of the 'Unique Pupil Number' (UPN) in 1999, followed by the 'Pupil Level Annual Schools Census' (PLASC) 3 years later. They provided the platform for the development of the 'Pupil Achievement Tracker' (PAT) and allowed data to be *matched* to individual pupils against a set of agreed background / context characteristics. It is this incorporation that paved the way for the current complex system of CVA measures (Ray 2006: 8). Further developments, driven by the so-called 'New Relationship with Schools' (DfES/Ofsted 2004), merged PandA and PAT into a new system, 'RAISEonline',[5] the aim of which was (and remains) to provide a common dataset for the dual purpose of accountability and school improvement; the bridge between the two aims being a structured process of school self-evaluation that provides the starting point for both Ofsted inspections and school improvement collaborations.

[5] RAISEonline—'RAISE' is an acronym for Reporting and Analysis for Improvement through School Self-Evaluation—was launched by the DCSF and Ofsted in January 2007 as a web-based portal providing access to a standard set of analyses of school-level and student-level data. It replaced the paper-based documents previously issued annually to schools; namely the Autumn Package and the Performance and Assessment (PANDA) reports. RAISEonline is part of Ofsted's New Relationship with Schools (NRwS), which aims to ensure that both schools and inspectors have access to the same set of data on which judgements of school performance can be made. RAISEonline also contains a suite of tools and analyses to use the same data for student and school target setting and school self-evaluation, which sets up a particular tension as the same data is used for both accountability and improvement purposes.

The adjustment of the value-added model to take greater account of pupil and school context clearly changed the picture of school performance as measured by the new model. Following the critique of value-added measures by Gorard (2006) we conducted our own analysis of published data for the 370 English secondary schools selected by the DfES (2005) for its KS2-4 CVA pilot, so that both the 'old VA' and the 'new CVA' scores were known for each school. We were thus able to investigate how the addition of pupil- and school-level contextual demographic variables in the CVA model affected the relationship between the raw unadjusted threshold performance measure (the proportion of students attaining 5+ A*–C GCSE grades) and the value-added measures. Our analysis showed that while 59% of the variance (r^2) in the basic KS2-4 ***VA*** scores of the schools could be explained[6] by the percentage of students attaining 5+A*-C GCSE grades (a lower value, although similar in magnitude to that obtained in Gorard's analysis), only 14% of the variance in KS2-4 ***CVA*** scores could be similarly attributed (see Figs. 1 and 2).

Despite the concerns we raise below around data and policy issues, in the era of initiatives like *Every Child Matters* in the UK and *No Child Left Behind* in the US, VA models like the one used in England take a clear step in a positive direction insofar as *every* student's data (potentially) makes a contribution to school performance (unlike threshold measures) and adjustment is made for both prior attainment and demographic context. As a result, the English models are currently being scrutinised by policy-makers in countries like Australia (Downes and Vindurampulle 2007) and Poland (Jakubowski 2008) as they prepare to develop their own school performance metrics.

3 Issue 1: Taking account of intake: The rationale for included variables

Research from the field of school effectiveness, conducted over more than three decades,[7] suggests that a major determinant of 'future success' is 'past attainment', so measures of prior attainment correctly form the basis for value-added modelling. In the KS2-4 CVA model, the variables that relate to prior attainment provide most of the model's explanatory power (typically around 50% of the variance), but there are important features to note in the way these are treated. At age 11, all students in maintained schools in England sit KS2 National Curriculum tests in English and mathematics (and previously also in science). The outcomes from the tests provide a measure of the academic attainment of each pupil in these core subjects by the end of the primary phase of their education. For reporting to parents and students the test outcomes are reported as broad National Curriculum *levels* and the expectation is that the majority of students should attain Level 4 or higher (Level 5 is the maximum possible) in each subject. For the purposes of providing a prior attainment measure, the levels are converted to a points score and the APS calculated and used as the

[6] Technically, 'degree of association' may be a more accurate term that 'variance explained' here, as the latter implies a causal relationship. For example, some commentators use '% variance explained' in the reverse direction with CVA as the explanatory variable for %5+A*–C. This is the opposite way round to Gorard (2006) who prompted our rebuttal analysis and who implied that as the old metric explained almost all the variance in the new metric, the new model was of little value.

[7] For a comprehensive review of the findings of school effectiveness research, see Sammons (2007).

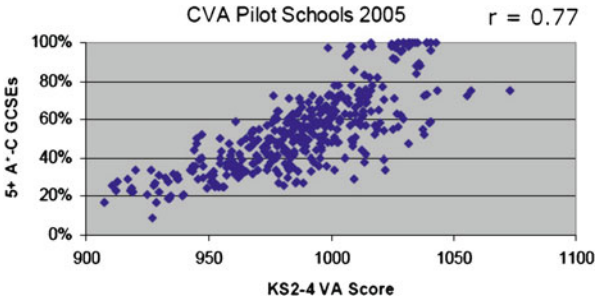

Fig. 1 Scatterplot and Pearson correlation of KS2-4 Value Added (*VA*) scores with percentage of students attaining 5 or more A*–C passes at GCSE for the 2005 KS2-4 CVA pilot schools, excluding special schools. (Source: DfES 2005)

main prior attainment variable in the model. Extra terms in the model are used to capture the *differences* between the APS and the English point score, and APS and the mathematics point score as these differences have been shown to be statistically significant in predicting the outcomes of students in GCSE examinations at age 16. A quadratic term (APS^2) is also included to reflect the fact that the relationship between KS4 outcomes and prior attainment at KS2 is non-linear.

Prior attainment measures are expressed as sub-levels ('fine grades'), rather than as crude levels like those used in earlier median VA models, though not going so far as to use actual percentages, which would make the model too complex. The model has two school-level variables that take the same value for every pupil in a given school: the mean *level* of school prior attainment and the *spread* of school prior attainment. The latter is included on the basis that schools with a narrow ability range 'more easily appear effective' and 'tend to have pupils with relatively high overall prior attainment' (Ray 2006: 43), but both prior attainment and its spread

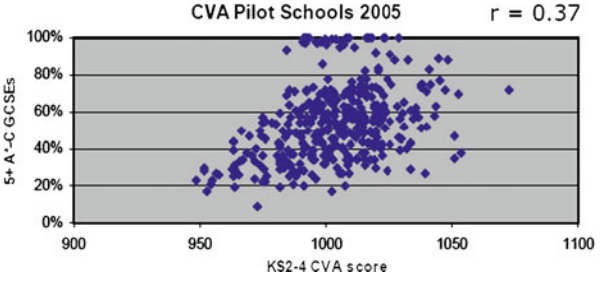

Fig. 2 Scatterplot and Pearson correlation of KS2-4 Contextualised Value Added (*CVA*) scores with percentage of students attaining 5 or more A*–C passes at GCSE for the 2005 KS2-4 CVA pilot schools, excluding special schools. (Source: DfES 2005). (An analysis of the effect of the additional contextualisation on the ranking of the pilot schools revealed that nearly half (46%) changed ranking bands (top 5%, top 20%, middle 50%, bottom 20%, bottom 5%) when rankings based on VA scores were compared to rankings based on CVA scores. Of the schools that moved bands slightly more moved up by at least one band than moved down. This was not the case for selective schools, however, with 75% of selective schools moving down by at least one band. Source DfES (2006) *Achievement and Attainment Tables: LA Conference 2006* (Available online at http://www.standards.dfes.gov.uk/performance/powerpoint/presentationLAv03.ppt?version=1) Accessed on 3rd December 2006)

only relate *to those pupils being modelled* (i.e. those in the GCSE cohort) and not to other pupils in the school at the same time, though the latter obviously and non-trivially impacts on learning.

In early VA models, pupil-level contextualising variables like socio-economic factors were not included, despite evidence of their importance from many years of school effectiveness research (Teddlie et al. 2000) because robust data on the socio-economic status of students was not then available (and possible because of the danger of lowering expectation). It was only with the advent of PLASC in 2002 that such data was available for all pupils in English maintained schools and enabled socio-economic context variables to be included in the model, though it is still not clear that the proxy measures used—typically, entitlement to free school meals—are suitably robust.

3.1 PLASC as the font of all knowledge

PLASC collects contextual data annually for use in the DCSF CVA model: on ethnicity, first language, gender, level of Special Educational Need (SEN), socio-economic indicators like free school meal (FSM) entitlement and deprivation measures associated with geographical location, and date of entry to school. This data is gathered and held by schools in order to make annual census returns to the DCSF, who then use the data to calculate CVA scores. Schools gather the data from individual pupil records held in their own data management systems, the majority of which will have been provided by parents in the form of paper returns to the school. Although the census return is made annually, considerable resources are invested in gathering and maintaining this pupil-level data.[8] The combined data from the annual census returns are stored in the National Pupil Database (NPD), which is maintained by the DCSF. The NPD is now a huge longitudinal dataset which contains attainment and contextual records for all pupils who have attended state maintained schools since 2002. It is available in various levels of detail to academic researchers and is made available to FFT to calculate pupil- and school-level value added data for local authorities and schools.[9]

Through PLASC data is gathered on all the main ethnic groups, plus ‘unclassified’ (<4%) for those individuals and schools who choose not to provide the information. The current range of ethnicity codes used in the CVA model is necessarily crude, especially for people of mixed background, and they ignore the ‘political’ subtleties of pupils claiming or denying their ethnicity, which go to the heart of whether pupils from some backgrounds feel culturally assimilated in British schools (and in British society generally). Provision exists to use an extended set of codes, but supporters of the model dismiss this possibility because extended codes are ‘not universally collected’ and would ‘add considerably to the complexity’ (Ray 2006: 22). Critics might retort that avoiding unnecessary complexity has not always been a priority for policy-makers; that the issue is more likely to be that changing the

[8] For a detailed account of how a typical school gathers information for its annual census return, see Rosina and Downs (2007).

[9] Students attending independent (i.e. private) schools contribute limited attainment data to the NPD, but no contextual data as these schools are not required to make annual census returns. CVA scores are therefore not calculated for independent schools.

codes would affect the consistency of the model over time, which *is* a legitimate concern (though more so for academics and policy-makers than for teachers). Generally, *parents* rather than pupils provide the data on ethnicity, which (it is claimed) makes it more accurate. Typically, less than 3% of pupils change ethnic category from year to year, so it is undoubtedly good quality data, but it does raise some subtle issues. Although the PLASC-informed CVA model takes account (in the form of coefficients and standard errors calculated in the normal way) of the relative size of the various effects and how significant they are compared to the size of the school effect, the published information[10] is not in a form that enables schools to use it *for improvement purposes*. It enables school and local authority staff to view fine-grained data and compare CVA scores for various subgroups, but it remains something of a 'black box' for many schools that operate in contexts significantly different from the national picture. For example, the proportion of students excluded from CVA calculations due to missing data is 5% nationally, but it is 11% in London and represents disproportionately those of non-white ethnicity with English as an additional language (Leckie 2007). Secondly, there is the issue of ownership. Have schools been empowered to capture, for example, the subtlety of pupils and families denying their own ethnicity in order to 'fit in'? If 'every child matters', what remediation is offered for the 3% who change ethnicity each year and who tells *their* story? Sometimes, what pupils think and say about their own ethnicities says as much about race in society as the statistics. What are schools supposed to *do* with ethnic (and similar) data in order to improve pupil outcomes, and are the assumed causalities immutable over time? If Asian girls born in August do better than Afro-Caribbean boys born in June, will this always be the case and how should teachers in the classroom act upon the information?

The data gathered from schools through the annual PLASC returns is used to generate the contextualising variables in the CVA model. Most pupil-level contextualising variables are simple binary (yes / no) answers to school census queries and are restricted to those for which national data is available, and in line with common practice in multivariate regression analysis a number of interaction terms are included in the model (for example, that between student ethnicity and FSM entitlement) where these have been shown to make a significant contribution to the model's explanatory power. The data on first language is one such binary code—'English' and 'Other than English'—with additional 'unknown' categories. It is non-problematic data, though it might be better to distinguish between languages other than English, especially where a different alphabet (or none) is involved. Data collection on gender is also fairly straightforward, but like ethnicity it is unclear what practitioners (as opposed to researchers) are supposed to *do* with the information.

The PLASC data on SEN is more complex as it covers a wide range of interrelated disabilities. The SEN Code of Practice offers three levels of response: 'School Action', where the teacher or coordinator provides the intervention; 'School Action Plus', where external help is sourced; and 'Statemented', where the child's educational needs are set out as entitlements by the local authority.[11] This is clear-cut in terms of action, but different local authorities *attribute the categories with varying*

[10] www.dcsf.gov.uk/performancetables.

[11] For the CVA model, the last two categories are treated as one.

degrees of liberality and this is not taken into account in the model; nor is the fact that some schools have their VA performance over time affected by factors like a local authority closing a neighbouring Special School.

In England, children whose parents receive certain welfare benefits have entitlement to free school meals, which is the main PLASC measure of social deprivation and family income. It is convenient but crude. As Ray (2006) notes, pupils who are *not* entitled to FSM vary considerably in their circumstance, just as there are degrees of deprivation within the FSM group itself. There is the additional problem that some parents who are entitled to FSM, out of embarrassment or for cultural reasons, elect *not* to claim it, and there are other families *not* entitled to FSM who are living in very straightened circumstances. As a consequence, the continued use of such a proxy for deprivation is a weakness and may undermine parallel school improvement initiatives based on the model.

Postcode data is the classifier of geographical location, which is also used to infer socio-economic status. Each student acquires an 'Income Deprivation Affecting Children Index' (IDACI) score, expressed as the proportion of children under sixteen in an area whose families are in receipt of income support. (A geographical area of measurement in the England has on average 1,500 people and subsumes a number of postcodes.) IDACI scores are calculated by another government department, not the DCSF, but it is a concern that the criteria used to define the categories *could* be manipulated to indicate success for particular government reforms. A non-governmental index—the 'Classification of Residential Neighbourhoods' (ACORN)[12]—is used instead by FFT in *its* models (Webber and Butler 2005), but whether the rationale for this is methodological or one of distrust is not known.

There are two variables capturing mobility: a simple binary one to cater for the effect of joining a school less than 2 years before the end of KS4; and another for pupils joining the school outside normal times of year (i.e. *not* in July, August or September of Years 7, 8 and 9). Supporters of the model claim that this keeps the model simple and intelligible, but again it is not clear that this is always the case and it is not clear that practitioners see it that way. Some variables, like 'change in FSM status from year to year', are excluded from the model—legitimately in our view—because they do not add enough to warrant the extra complexity, but other variables, like the percentage of pupils from minority ethnic communities, are excluded even though they appear *prima facia* to be important.

Finally, PLASC data is collected on pupil absences and exclusions, but this data is not used in the CVA model because although it would improve its explanatory power, it is thought that 'schools should to some extent be responsible' for the factors employed (Ray 2006: 21). Here we see the recurring tensions nicely encapsulated: it is unclear whether the purpose of PLASC and its dependent CVA model is prediction, accountability or improvement. Certainly, it is proper that the model should only contain the factors / inputs that are outside the school's control so that what is left is the school effect, but it is arguable whether factors such as SEN and truancy are within or without the control of the school. The inclusion of certain factors enhances the utility of CVA for improvement purposes, but the 'one size fits

[12] ACORN was developed by *CACI International*, formally the *Consolidated Analysis Centers Inc*. For more information on this geo-demographic measure, see http://www.caci.co.uk/msd.html.

all' approach creates tensions for accountability purposes (and vice versa) and leads to the creation of perverse incentives whereby schools can manipulate the metrics, particularly in the area of SEN and ethnicity, to get better scores (Ray 2006). This *could* be overcome by having counterbalancing incentives elsewhere, but it is not clear that policy-makers have covered all the bases in this respect; that for every manipulation to raise false CVA scores there are equal and opposite incentives for schools *not* to do so.

4 Issue 2: Issues related to the data and statistical models used to generate VA scores

CVA models, like VA models before them, use national data from KS tests and various pupil and postcode characteristics to capture a school's intake. Every pupil is tracked through their UPNs with only a tiny amount of wastage in the data (although up to KS4, only data from the core subjects—English, Mathematics and Science—is collected). National Curriculum tests at the end of the Key Stages were introduced in England to measure attainment in certain subjects, not to calculate added value. Originally, national test scores were reported as broad 'National Curriculum Levels' with particular attention paid to moderating the examination scripts *of borderline students*, but for the purpose of calculating CVA scores, a finely graded system is required in which test scores are reported to the nearest *tenth of a* level. Whether the current system in England is robust enough to support such fine divisions in grading is the question. While the measurement error introduced by the subjectivity of the marking process will balance out during the aggregation that produces the complete national dataset, the current practice of allocating all National Curriculum tests for a single school *to a single examiner* may result in bias in the test scores for an individual school. This creates issues for the use of CVA for both accountability and self-evaluation purposes. It *would* be possible to use the technology of electronic marking to assign randomly scanned scripts within the bank of test markers to ameliorate any potential bias, but problems with the recent introduction of electronic administration and marking for the 2008 National Curriculum tests suggests this may be no small task (BBC 2008).

Secondly, the scoring system itself is a concern. Pupils taking KS2 tests are given marks in English, mathematics and science, which are then converted into one of three crude overall levels: Level 4 is designated as the average expected attainment for age 11; Level 5 is above average; Level 3 is below average. Obviously, there is a ceiling effect for bright pupils (Tymms and Dean 2004) who cannot score above a '5' though they may be operating well above that level, and since approximately one-third of pupils get the top grade, this may have a significant impact.[13] From a school's point of view, each of the three levels is assigned a points value according

[13] After the experience of the 2005 CVA pilot, an adjustment was made to the 2006 scores to account for ceiling and floor effects in the predicted CVA calculations to bring them into line with the actual capped point scores observed at the extremes of the data. That said, there is clearly a limit to the capped (i.e. best eight) GCSE point score (8×58=464) so that a student with the same prior attainment and the same capped points score as another, but with a higher total GCSE point score will receive no extra 'credit' in his or her value added score. It is easier to make such adjustments for ceiling and floor effects in multi-level models than in median VA models.

to a formula, and an average points score is then calculated across the three subjects. However, despite using a fine-grade system in the measure of prior attainment,[14] pupils who are scoring (or assumed to be scoring) below Level 3 receive the same number of points regardless of whether they were *just* below the Level 3 threshold or whether they had SEN and were not entered for the tests, so there is also a 'floor effect' to the data. The three levels are designed to be aligned over time with levels at other Key Stages—in other words, that Level 4 (say) means the same thing in any subject at any Key Stage—but as Ray (2006: 18) points out, there may be some 'misalignment' (Massey et al. 2003).[15] It may be true that equivalence between KS levels is not essential for a value-added system as long as all schools take the same tests, but this is not true between years or between levels. Level 5 in Mathematics in 2001 might not represent the same learning as the same level in the same subject in 2006, and a Level 3 might not represent the same level of underperformance relative to Level 4 in 2001 as it did in 2006. For school effectiveness research, this diminishes the value of any comparison of a school's CVA scores over time, and for practitioners it suggests that a school can only be as good as its previous year's examination results. Supporters of the current system claim that normalising KS results so that they have the same distribution in any given year would lose the public meaning of 'levels' in terms of the learning they represent, but the same logic has not been applied to examinations like GCSE and (university entrance) A-levels, where equivalence across qualifications, subjects and courses is almost impossible to fathom and has resulted in widespread system-playing in many schools. If the priority for policy-makers is a system that is easy for classroom teachers and parents to understand, even if that means sacrificing some robustness, then that should also apply to CVA measurement.

As discussed above, the scores generated by CVA models are calculated from the differences between the predicted and actual attainment of pupils, taking into account a wide range of contextualising factors: ethnicity, speaking English as an additional language, gender, age, level of special educational need, entitlement to free school meals and income deprivation, late entrance to school, and being 'in care'. In England, for the CVA model used for pupils between the ages of eleven (KS2) and sixteen (KS4), this entails trying to predict results for up to eleven GCSE subjects per student from prior attainment in three (English, mathematics and science) achieved 5 years previously. The calculation uses a multilevel model (MLM)—an excellent development in itself—to take account of the fact that pupils are grouped by school and are not independent of each other; in other words, that the data is hierarchical. In each CVA model, the residual variance is partitioned into two

[14] The CVA models used by both DCSF and FFT employ a fine grade system. For example, when calculating KS2-4 CVA, the raw scores achieved in each National Curriculum test are used to calculate a decimalised level of point score. A student attaining Level 4 at KS2 under the old VA system would have been awarded a point score of 27. Using fine grades under CVA, those students only just scoring a level four would get a point score of 24.0 whereas those right at the top of the Level 4 band get a point score of 29.9, which provides a continuous scale for APS. For more details of the specific steps in the calculation and build of the CVA model see DCSF 2008a.

[15] In the May 2007 edition of '*Inspection Matters*', Ofsted stressed to school inspectors that CVA was a *relative* rather than an *absolute* measure of performance and that care should be taken when interpreting trends in CVA scores, as MLM coefficients are recalculated each year and adjustments made to the factors included in the model.

levels: pupil level ('Level 1') and school level ('Level 2'), which are the model's random effects. It is possible to include other levels in the model—class groups within a school, for example, or within a local authority for groups of schools—but this has not been done. The MLM also applies a 'shrinkage' factor to residuals as part of the process of calculating school CVA scores, which produces more robust estimates of the statistical significance of the scores for schools with small cohorts. This has the effect of causing the scores for small schools, and for particularly high- or low-scoring schools, to be moved closer to the national mean, so making it less likely that extreme scores are recorded.

Multilevel modelling, unlike Ordinary Least Squares (OLS) methods, offers a more complex set of options to take account of the data-structuring fact that pupils are grouped by school. One of the advantages is that it produces more robust estimates of the standard errors for factors in the model (whereas OLS methods tend to *under*estimate them), which means that judging whether or not a factor is statistically significant is more rigorous. Another difference between OLS and MLM is in the way the latter 'shrinks' the VA estimates, depending (in part) on the size of the school. Application of the shrinkage factor means that the CVA score is reduced to a percentage of its raw size, closer to the mean,[16] which process has been described by Kreft (1996) as akin to small schools 'borrowing strength' from data from larger schools. The resulting shrunken scores have narrower confidence intervals and these are given in performance tables (for KS2-4). As mentioned above, shrinkage prevents schools at the extremes—those with residuals which suggest that they are either very effective or very *in*effective—from registering a very high or a very low CVA score. Supporters say that this is non-problematic because the raw residuals for small schools are anyway known to be poor estimates of effectiveness from 1 year to the next, but Fitz-Gibbon (1991: 19), quoting Raudenbush, one of the early developers of MLM in the field of school effectiveness, suggests that shrinkage causes scores to be pulled in a 'socially expected direction, demonstrating a kind of statistical self-fulfilling prophecy'.

Notwithstanding the decision to opt for complexities such as shrinkage with the introduction of the multilevel model, strange 'reversions' occur when the system goes practical. Ofsted, for example, now provides inspectors with the means to *un*-shrink the data to see what the raw residuals look like, so that one wonders at using a system to shrink the data in the first place. If inspectors can judge from the raw data and from their own impressions the extent to which the raw residual is 'an accurate reflection' of a school's effectiveness, why all the complexity at the practitioner level? It suggests that CVA measurement is more aligned to the agenda of public accountability and performance tables than to the critical process of self-evaluation as espoused by the government's 'New Relationship with Schools' (DfES/Ofsted 2004). This is compounded by the fact that the application of shrinkage factors causes problems in the calculation of CVA scores *for sub-groups* of pupils that lie at the heart of the self-evaluation tools used in RAISEonline. A recent attempt by Hillingdon Borough Council in London, for example, to assist its schools in

[16] Generally, OLS has a problem dealing with small cohorts so that a small school's score can only be given with a wide confidence interval. With MLM, national data is used to modify the estimate when information on the school is limited because of size.

interpreting RAISEonline outputs illustrates the problem (Thomson 2007: 41). In one sample output, the CVA score for 182 matching pupils in a cohort was 978.4; the scores for the 170 students *with* English as their first language was 978.7; while the 12 students who did *not* have English as their first language had a CVA score of 987.3. The scores for both subgroups were therefore higher than the average score for the combined cohort, which is confusing information for managers and teachers who would intuitively expect to see the scores for subgroups distributed around the mean for all students. In this case, the application of shrinkage has pulled all three residuals closer to the mean (1000), but it had a greater effect on the sub-group residuals due to their smaller sizes.[17] The point here is *not* that there is any theoretical flaw in the modelling or that the system is unfair or unjustified, but that it is important at the practitioner level that schools be briefed extensively on issues like shrinkage to prevent imposing a barrier of expertise around the data.

5 Issue 3: Policy issues related to development of VA measures

The Value Added National Project of 1997, conducted by researchers at the University of Durham (Fitz-Gibbon 1997)—an extensive study of value-added in the UK—advised that simple and easy-to-understand measures should be used in preference to slightly more robust, but much more complex, multilevel models.[18] While the logic of this approach was considered sensible at the time, 'many of the experts consulted by the DfES' favoured complexity and so 'it was decided to move ahead on this basis' (Ray 2006: 49). While multilevel models have admirably corrected for grouping effects, there has been little by way of debate as to whether the complexity added by such advancements is justified by the significance, in the wider sense, of the measures to practitioners. Despite opposition from some quarters (e.g. Gorard 2006), there is little doubting the benefit of the modelling to school effectiveness research, but if, for example, what is being measured—how well pupils do at examinations and how much better or worse they do as they get older—is accepted as being only a small part of the *education* they receive in school, and if 'the school effect' is anyway accepted as being relatively small, one must ask

[17] In the light of such 'shrinkage discrepancies', the current advice given by Hillingdon Local Authority to its schools is to avoid using *RAISEonline* for self-evaluation involving student sub-groups, but to use FFT analyses instead (Thomson 2007: 76).

[18] At the time of the *Value Added National Project*, practitioners supported the simple approach but accepted that performance tables should contain some measures to account for different pupil intakes. As a result, the DfEE (1998a) trialled a measure of KS3-4 added value that compared pupil attainment at KS4 (for the 1998 cohort) with the national median KS4 attainment for pupils from the previous year at each level of KS3 prior attainment. Scores for schools were then calculated as the average of these differences (Critchlow and Coe 2003). This 'median method' was easy for parents to understand and for professionals to interpret, and was consistent with how national data was presented to schools at the time (DfEE 1998b). The response to the trial was positive and although some commentators suggested that the absence of *pre*-KS3 measures was problematic, there was widespread and patient acceptance that these would be added in due course. In deference to the reservations expressed, the (then) DfES agreed not to publish school VA scores until the full compulsory secondary school range, from KS2 to KS4, could be covered by the data streams. This first occurred in 2001 when a pilot was done of KS2-3 value-added to add to the existing measures from KS3-4, and in 2002 this 'double suite' of VA scores was published for all secondary schools. Two years later, VA measures for pupil attainment from KS2 to KS4 were added.

whether the obfuscation that results from the complexity of ever more accurate measures is worthwhile when ever fewer people can understand and interpret the results. *Ad absurdum*, if only a handful of academics understand the models to the extent of being able to challenge them, there is little use in them for professionals whose very essence lies in understanding challenges to practice and accommodating change.

5.1 Taking account of context

The first PLASC in 2002, linking individual pupils to their achievements through UPNs, afforded the opportunity to include contextual data alongside the prior attainment of pupils. Other organisations, like FFT and London Families (DfES 2006), developed their own CVA models alongside the DfES one. The government was aware of the need to retain the confidence of stakeholders as the models proliferated (Miliband 2004), but slowly the voices in favour of greater complexity began to command the stage. It is difficult to see how policy-makers could have withstood the advice of pro-complexity advocates in favour of an easier-to-understand practitioner approach. The more obtuse the arguments in favour of complexity, the less anyone could contest them, least of all policy-makers who were neither statisticians nor practitioners. Somewhere in the excitement, the importance of access for practitioners was mislaid or underestimated. In 2005, a CVA multilevel model for secondary schools was piloted using PLASC data, which Ofsted used in its PandA reports and (with some amendment) still uses today as the basis for RAISEonline. However, even supporters of complex modelling acknowledge that there is a difficulty 'maintaining continuity' (Ray 2006) as practitioners move to multilevel models from simpler versions, and discussion is ongoing about which ones are best suited to certain situations and how best to present and interpret the results. It has been suggested that these problems are largely transitional, but it seems to be more than this. A point seems to have been reached where some contextual factors are downplayed because they add too much complexity, but others are included simply because they are easy to measure. For example, FSM entitlement is still used in CVA models though the accuracy and appropriateness of its use as a proxy for economic disadvantage is questionable. Month of birth is also taken into account in CVA because by including as many factors as possible outside the school's control, the residual / difference between the model and the pupil data comes closer theoretically to the school effect, but month of birth is of little use to practitioners if the measures are to be used for school improvement purposes. If CVA is to be a *predictor* of future attainment, then every factor should be used in the model; if CVA is to be used for *school improvement* purposes or to allocate funding, there is little point including factors that cannot be changed;[19] if CVA is to be used for *accountability* purposes, the model must above all else be understandable to the extent that it does not require 'outside' expertise; and if CVA is to be used for a *combination* of the above then decisions need to be made that balance the competing claims of statistical robustness, usability and accessibility. Generally, the school effectiveness research community would welcome such a differentiation; it is not in anyone's interest to perpetuate the inappropriate use of important data.

[19] Except insofar as the remaining variance is then closer to what the school adds.

There has been little conceptual debate about the use of CVA as a predictor of attainment beyond comparing what *is* achieved with what *might have been* achieved if nothing else had changed other than what the school did or failed to do. Whether this paucity of debate is due to the fact that this assumption—that everything remains constant except what the school does—is unrealistic, or to the unwillingness of policy-makers to contest the desirability of pretending to 'know' what students are capable of becoming, is unclear, but the rush to complexity by the UK government continues unabated. Likewise with the inclination to 'one-size fits-all' analyses. RAISEonline is an interface between practitioners and the 'black box' of CVA prosody. Certainly it provides schools with a better range of data and outcome measures for pupils, but is it a helpful imposition on *all* schools to use a *common* method of analysis and to have to judge performance against national patterns without fully understanding how the calculations are being made and under what assumptions? The fact that data from RAISEonline is also used for self-evaluation merely creates the *illusion* of ownership, and the availability to Ofsted of the analysed data means in effect that schools are now controlled not only in *what* they do (via the National Curriculum) and in what data they collect, but in *how* they judge what they do. This in turn, it could be argued, has a de-professionalising effect on teaching and headship. The simple fact that teachers use data (Kirkup et al. 2005) should not be an end in itself, and it is certainly not enough to constitute a long-term strategy for school improvement. In particular, it is no small task to tease out the contribution made *by individual teachers*, especially in secondary schools where students change teaching groups within the school year and have had a range of different teachers for any given subject across the years of a Key Stage. Still more pupils will have had access to extra tutoring. Yet few measures have incorporated the 'intermediate' level of the teacher into multilevel models, though a number of studies have suggested that the magnitude of the teacher-level effect may be of the same order of magnitude as the school effect in terms of the variance partitioned to each level (Luyten 2003; Sharp and Croxford 2003; Sharp 2006). While schools can justifiably be considered responsible for much of the variation at class level, such findings call into question the usefulness of a single value-added measure to inform parental choice, the longstanding rationale behind them. While parents now have the facility, to varying extents around England, of being able to choose their children's schools, nowhere do they have the facility to select their children's teachers.

5.2 Issues of comparison and compatibility

There is an additional concern now in the UK that the organisations that run the testing system are not independent enough of those that set policy and inspect schools, particularly when higher scores are trumpeted by policy-makers as proof that certain policies are 'working'. Statistical analysis *does* indicate that the examination system in England is testing what it claims to be testing, but this is not the same as claiming that the tests are testing what they *should* be testing, and there is no shortage of evidence that teachers are ignoring broader educational outcomes (Volante 2004; James 2006). Given this fact and the widely held belief that pupils in England are over-tested, it is unclear why sampling is not being used throughout the system, as it is in the Foundation Stage for five-year olds. Why, if the tests are valid and reliable, is it

necessary to test *every* child from *every* postcode and from *every* ethnic background in order to evaluate how policies are impacting on children? It should be a simple matter to arrange an appropriate sampling frame and thereby make huge savings in terms of cost and interruption. And there are other anomalies in addition. Firstly, test data for KS1, KS2 and until recently KS3 is only collected for the core subjects—English, Mathematics and Science—which fact itself is likely to skew both teaching and the data, particularly at KS3 when secondary school pupils have typically been studying eleven subjects for 3 years without having their proficiency in eight of them gauged by national tests. Admittedly, the cost of GCSE-type assessment at earlier key stages would be cost-prohibitive and disruptive, especially if it contained in-house assessment, but one solution would be to sample externally moderated, but internally assessed, formal school examinations (rather than coursework, which is fast losing credibility). The fact that the government does not address this situation is, according to Tymms and Dean (2004), evidence that it is confusing 'robustness' with 'bias'. Generally, the greater the range of evidence used to make an assessment the more robust its claims, and since classroom teachers are best placed to offer this, the government's reservation about sampling must therefore be one of bias; in other words, that not all teachers will be equally exacting in their tests.

6 Conclusion

Value-added measures represent a considerable improvement on threshold measures, both in terms of *what* is being measured (progress adjusted for prior attainment rather than raw outcomes) and *who* is being measured (*all* pupils rather than just those who cross an arbitrary threshold). That notwithstanding, it may be that the model in England falls foul of trying to be all things to all people. Despite its complexity, even for an academic audience, it represents in some ways an inappropriate *over-simplification* of the nature of school performance. If pupil attainment could be measured by academic outcomes alone, and across a narrow range of public examinations, school CVA scores would not capture the differential effectiveness of schools across the range of prior attainment and across the various sub-groups.[20] There are shortcomings for practitioners too, in terms of timing and accuracy. The measures are fixed on *provisional* contextual data and schools are not given an opportunity, subsequent to checking data initially, to make amendments. The DCSF says that giving schools such an opportunity would affect the timescale of publication and might necessitate a second round of checking, but the pressing desire to publish means that the final scores and their confidence intervals are not in fact direct outputs from the multilevel model (Ray 2006: 51), but are calculated by outside contractors hired for the purpose. And when school value-added scores *are* published, they have artificial ceilings imposed on them so that they are not greater than the theoretical maximum, in response to which Schagen (2006) has suggested standardising residuals at each prior attainment point and stretching small differences

[20] The measures *could* be used to reveal differential effectiveness—by presenting each variable against the national expected performance for that category—but that would mean further complicating the model for a public audience (Gorard 2006).

in the scale for 'extreme' pupils so that they have more chance of affecting the outcome. This is a similar approach to the one used by FFT to deal with non-linearity and ceiling effects; namely, to calculate a predicted KS4 outcome from KS2 input and then feed this back into the model as an input.[21]

The model meets its public audience when the shrunken residuals are used for ranking schools, and this causes its own problems (Kreft 1996; Gorard 2007; Hutchison and Schagen 2008). Admittedly, the DCSF *itself* does not rank schools in its annual published performance tables, but it is aware that the shrunken scores it puts into the public domain are ranked by national and local media. This practice has led to calls for the inclusion of (95%) confidence intervals in the performance tables (Goldstein 2007) in the hope that their publication (or that of some other marker of statistical significance) will be taken up by the media in its league tables. While the DCSF has risen to the challenge, the media has yet to reciprocate,[22] so that in some ways the optimal public use of its measures is hindered by the government's own drive towards greater complexity. It was not always thus. When the UK government first published VA scores for primary schools in December 2003, the then Schools Standards Minister, David Miliband, stated:

> 'We have always said that we will listen to the views of heads, teachers and parents about how performance tables can provide a more comprehensive and rounded picture of school performance. Including value added information does just that. It shows the rates of progress that children make between 7 and 11 in different schools.' (GNN 2003)

It may be that the Minister's use of the term 'rounded' carried more meaning than was ascribed to it at the time—the danger of resorting to a 'one-number-fits-all' approach to measuring the performance of an individual school is that it presents an overly simplified view of the school's effectiveness that belies the complexity of the metric—but does less to inform parental choice and inspection judgements than its supporters claim. Whether or not published CVA scores are accompanied by confidence intervals, and whether or not they are published as true residuals, they suggest a degree of precision in the measurement of school performance that is not justified. And despite their complexity, the measures fail to respond adequately to competing legitimate demands: from the public for *interpretability*; from teachers for *usefulness*; and from policy-makers for *accountability*.[23]

[21] Interestingly, Schagen, who is certainly no stranger to MLM, chooses to use OLS to calculate the predicted GCSE scores in the 2006 paper referenced here.

[22] Goldstein, whose work has done much to shape the development of CVA, has expressed his frustration at the absence of confidence intervals in the league tables published by the major UK newspapers and the media. While he praised the BBC website for giving a balanced overview of the issues relating to the use of CVA scores in league tables (Eason 2007), the same website also contained interactive tables of school CVA and 'raw' threshold measures against which schools could be sorted against any of the measures at the click of a mouse, without a confidence interval.

[23] The newly proposed 'School Report Card' (DCSF 2008b), which looks set to combine data from a wide range of school performance metrics *including CVA*, may fall foul of these same issues if the DCSF holds fast to its determination to provide a single overall grade for each school based on a (yet to be revealed) weighting formula that schools suspect will be changed from year to year to suit the prevailing political agenda.

References

Aitkin, M., & Longford, N. (1986). Statistical issues in school effectiveness research. *Journal of the Royal Statistical Society, Series a, 149*, 1–42.

BBC (2008). News. *Delays hit pupils' test results*. Available online at: http://news.bbc.co.uk/1/hi/education/7489510.stm. Accessed on 15/02/2009.

Critchlow, J., & Coe, R. (2003). *Serious flaws arising from the use of the median in calculating value added measures for School Performance Tables in England*, Paper presented to the 29th International Association for Educational Assessment Annual Conference, October 2003.

DCSF (2008a). *Technical guide to CVA—2007/08 model*. Available online at: www.dcsf.gov.uk/performancetables/schools_08/documents.shtml. Accessed on 3/6/2009].

DCSF. (2008b). *A school report card: Consultation document*. Nottingham: Department for Childern, Schools and Families.

DfEE. (1998a). *1998 value added pilot: Supplement to the secondary school performance tables*. London: Department for Education and Employment.

DfEE. (1998b). *The autumn package*. London: Department for Education and Employment.

DfES/Ofsted. (2004). *A new relationship with schools: Improving performance through school self-evaluation*. Nottingham: DfES.

DfES (2005). *2005 key stage 4 Contextual Value Added (CVA) pilot*. Available online at: http://www.dcsf.gov.uk/performancetables/pilotks4_05.shtml. Accessed on 12/11/2006.

DfES. (2006). *Families of schools: May 2006 secondary schools*. London: Department for Education and Skills.

Downes, D., & Vindurampulle, O. (2007). *Value-added measures for school improvement*. Melbourne: Department of Education and Early Childhood Development.

Eason, G. (2007). *How the new CVA scoring works*, BBC News. Available online at: http://news.bbc.co.uk/1/hi/education/6251587.stm. Accessed on 24/08/2007.

Fitz-Gibbon, C. T. (1985). A-level results in comprehensive schools: the COMBSE project. *Oxford Review of Education, 11*(1), 43–58.

Fitz-Gibbon, C. T. (1991). Multilevel modelling in an indicator system. In S. W. Raudenbush & J. D. Willms (Eds.), *School, classrooms and pupils: International studies of schooling from a multilevel perspective*. San Diego: Academic.

Fitz-Gibbon, C. T. (1997). *The value added national project final report: Feasibility studies for a national system of value-added indicators*. London: School Curriculum and Assessment Authority.

GNN (Government News Network) (2003). *Value added results show more rounded picture of primary schools' progress—Miliband*. Available online at: http://www.gnn.gov.uk/content/detail.asp?NavigatedFromSearch=True&NewsAreaID=2&ReleaseID=101891. Accessed 06/09/2007.

Goldstein, H. (2007). *Evidence and education policy—some reflections and allegations*, Paper presented to the RSS conference in York, July 2007. Available online at: www.cmm.bristol.ac.uk/team/HG_Personal/Evidence%20and%20education%20policy.pdf. Accessed 24/08/2007.

Gorard, S. (2006). Value-added is of little value. *Journal of Education Policy, 21*(2), 235–243.

Gorard, S. (2007). The dubious benefits of multi-level modelling. *International Journal of Research & Method in Education, 30*(2), 221–236.

Gray, J., Jesson, D., & Jones, B. (1986). The search for a fairer way of comparing schools' examination results. *Research Papers in Education, 1*(2), 91–122.

Hutchison, D., & Schagen, I. (2008). Concorde and discord: the art of multilevel modelling. *International Journal of Research & Method in Education, 31*(1), 11–18.

Jakubowski, M. (2008). *Implementing value-added models of school assessment*. Florence: European University Institute.

James, M. (2006). University of London Institute of Education, reported on the BBC, 9 August 2006. Available online at: http://news.bbc.co.uk/1/hi/education/4777737.stm. Accessed 13/09/2007.

Kirkup, C., Sizmur, J., Sturman, L., & Lewis, K. (2005). *Schools' use of data in teaching and learning, DfES research report 671*. London: Department for Education and Skills.

Kreft, I. G. (1996). *Are multilevel techniques necessary? An overview including simulation studies, cited in: C.T. Fitz-Gibbon (1997) The value added national project final report: Feasibility studies for a national system of value-added indicators*. London: School Curriculum and Assessment Authority.

Leckie, G. (2007). *Missing data and school assessment measures*, presentation to the PLASC/National Pupil Database Users Group (PLUG), University of Bristol, 17 January 2007. Available online at www.bris.ac.uk/cmpo/plug/workshops/index.html.

Luyten, H. (2003). The size of school effects compared to teacher effects: an overview of the research literature. *School Effectiveness and School Improvement, 14*(1), 31–35.

Massey, A., Green, S., Dexter, T., & Hammet, L. (2003). *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001. Final Report to the QCA of the Comparability Over Time Project*. London: QCA.

Miliband, D. (2004). *Personalised learning: Building a new relationship with schools*, Speech by David Miliband, Minister Of State For School Standards, North of England Education Conference, Belfast, 8th January 2004.

Ofsted (2007). *Inspection matters*, Issue 14, May 2007.

Ray, A. (2006). *School value added measures in England: A paper for the OECD project on the development of value-added models in education systems*. London: Department for Education and Skills.

Rosina, H., & Downs, M. (2007). *Collecting data for a school*, presentation to the PLASC/National Pupil Database Users Group (PLUG), University of Bristol, 17 January 2007. Available online at www.bris.ac.uk/cmpo/plug/workshops/index.html.

Sammons, P. (2007). *School effectiveness and equity: Making connections*. Reading: CfBT Education Trust.

Schagen, I. (2006). The use of standardized residuals to derive value-added measures of school performance. *Educational Studies, 32*(2), 119–132.

Sharp, S. (2006). Assessing value-added in the first year of schooling: some results and methodological considerations. *School Effectiveness and School Improvement, 17*(3), 329–346.

Sharp, S., & Croxford, L. (2003). Literacy in the first year of schooling: a multilevel analysis. *School Effectiveness and School Improvement, 14*(2), 213–231.

Teddlie, C., Stringfield, S., & Reynolds, D. (2000). Context issues within school effectiveness research. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research*. London: Falmer.

Thomson, D. (2007). *Using RAISEonline in self-evaluation. Understanding what the inspectors will see. (Secondary)*. Available online at: www.egfl.org.uk/export/sites/egfl/categories/admin/data/_docs/raise_online/raiseSec.pdf. Accessed 30/08/2007.

Tymms, P., & Dean, C. (2004). *Value added in the primary school league tables, a report for the National Association of Head Teachers*. Durham: CEM Centre, University of Durham.

Volante, L. (2004). Teaching to the test: what every educator and policy-maker should know, *Canadian Journal of Educational Administration and Policy,* Issue 35. Available online at: http://www.umanitoba.ca/publications/cjeap/articles/volante.html. Accessed 17/11/07.

Webber, R., & Butler, T. (2005). *Classifying pupils by where they live: How well does this predict variations in their GCSE results? CASA Working Paper Number 99*. London: Centre for Advanced Spatial Studies, University College London.

Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: Falmer.