Value-Added in the Primary School League Tables

A Report for the National Association of Head Teachers

Prof. Peter Tymms and Dr. Colin Dean

CEM Centre, University of Durham

May 2004

Contact Details – NAHT

Arthur De Caux & Christine Girdler
Education Department
National Association of Head Teachers
1 Heath Square
Boltro Road
Haywards Heath
West Sussex
RH16 1BL
United Kingdom

Phone: +44 (0) 1444 472472
Fax: +44 (0) 1444 472473
Web: www.naht.org.uk
Email: info@naht.org.uk

Contact Details – CEM Centre

Prof. Peter Tymms & Dr. Colin Dean
Curriculum, Evaluation and Management Centre
University of Durham
Mountjoy Research Centre 4
Stockton Road
Durham
DH1 3UZ
United Kingdom

Phone: +44 (0) 191 334 4185
Fax: +44 (0) 191 334 4180
Web: www.cemcentre.org
Email: info@cem.dur.ac.uk

Contents

NAHT Introductory Remarks

From the outset, the NAHT has remained opposed to the publication of performance tables on the grounds that they are unfair, misleading and have a strongly negative effect on assessment, on the breadth and balance of the curriculum and on teacher morale.

In particular the Association cannot understand how parental choice is being properly informed when pupils with statements, units containing pupils with significant learning difficulties and absent pupils are counted in the calculation of percentages for the tables when they have no positive result to contribute to the threshold scores.

Whilst we accept that a value added system is likely to be less unfair, we remain very concerned at the way it is being calculated by the Department for inclusion in the performance tables.

Value added information, sensibly interpreted, is very helpful to teachers and school leaders in assessing performance, areas of strength and weakness and as an aid to school improvement.

Schools, with guidance and training, can interpret, with understanding and caution, the statistical reliability of their data. However, value added information loses credibility when it is published, especially in high stake performance tables, taking little or no account of size of cohort, the reliability of input and output measures, SEN small steps or the 'ceiling effect' on high achieving pupils, whereby it is virtually impossible for anyone with a very high input score to achieve a positive value added score.

The NAHT continues to press for abandoning of all forms of performance tables, as is now the case in Scotland, Wales and Northern Ireland. But if they are to remain in England, then they have to be in the form of the 'least flawed' value added model, accompanied by suitable words of caution.

To help create a better understanding of the issues involved and to develop a valid model to help the self-assessment by schools, the Association has entered into this research project with Durham University.

Arthur de Caux,
Senior Assistant Secretary, Education,
National Association of Head Teachers,
April 2004.

Summary

The DfES published value-added scores in the school performance tables for primary schools in England nationally for the first time in 2003. These were intended to allow schools to demonstrate the progress made by children during Key Stage 2.

Through discussion with head teachers and statistical analysis, we have found a number of issues concerning the way that the results have been calculated and presented which need urgent attention. In particular:

- The small size of many primary schools will result in wild fluctuations in the official statistics from year to year. Data are published for schools with as few as 11 pupils in a year group and this means that, even when the quality of a school's educational provision remains the same, its indicators will give the impression of change.

- There are issues with the quality of the end of Key Stage 1 data used to calculate value-added scores. Internal marking and the possibility of differences between infant schools and all through primary schools make the validity of comparisons between the different schools questionable.

- Value-added scores are assigned to individual schools on the basis of pupils who were attending that school when the Key Stage 2 tests were taken. But this does not take account of pupil turnover and the fact that, for many schools, a high proportion of their pupils were not on their school roll for all of the preceding 4 years.

- The chosen statistical methodology generates surprisingly large errors of measurement and biases the published results: schools with high-achieving pupils at the end of Key Stage 1 are prevented from achieving high value-added scores at the end of Key Stage 2.

- As published, the results do not compare like with like. It is unfair to hold schools to account based on comparisons between schools in affluent areas and those in deprived areas. Contextual factors may have a significant effect on a school's success. These include differences in parental support, neighbourhood influences, peer-group effects and the proportion of children whose first language is not English.

We conclude that, although value-added information is an essential tool for professionals, the publishing of value-added indicators in their current form is misleading and should be discontinued.

Prof. Peter Tymms & Dr. Colin Dean
Curriculum, Evaluation and Management Centre,
University of Durham,
May 2004.

Overview

Primary school pupils in England sit National Curriculum tests and carry out tasks at the end of Key Stage 1 (KS1), aged 6 or 7, and at the end of Key Stage 2 (KS2), aged 10 or 11. Each year the Department for Education and Skills (DfES) publishes tables showing the results achieved by schools at KS2. For the first time, nationally, in 2003, the DfES also published a KS1 to KS2 value-added measure for schools, with the aim of showing how much progress pupils make at each school, with the intention of allowing fairer comparisons between schools whose pupils arrive with different needs and abilities. Value-added measures had been published previously for secondary schools, but this was the first time similar tables were made available for primary schools across the whole of England.

There are undoubted benefits to teachers and other educational professionals in being able to measure how much progress individual pupils and whole schools make. As one primary head teacher comments:

"[The introduction of value-added rather than just KS2 test results] is definitely a move forwards." (H1[1])

However, there are several very serious issues in calculating and using primary school value-added measures, both generally and specifically, in relation to the DfES's chosen methodology. Problems also arise from the way in which the results are currently presented and published. Another head teacher observes:

"We have 30 moderate learning difficulty pupils in a primary school, in two special needs units, and certainly the introduction of the value-added has meant that those children's attainments are taken into account in a more realistic way than the threshold manner. It still doesn't totally resolve the situation of having an accurate picture of our children and the way in which they have developed through school, and the successes and achievements they make." (H5)

In this report, we have attempted to identify and discuss these issues, drawing on both statistical analyses and conversations with primary head teachers. We conclude that there are fundamental flaws in the current methodology and limitations in the data it uses, which not only introduce random errors[2], but also lead to biases[1] against certain categories of pupils and don't provide a level playing field against which to compare different schools. Our investigations also lead us to question the criteria used to decide whether a particular school's results should be published at all, and to doubt the validity of comparing schools without considering the different contexts within which they find themselves.

We have also included some specific suggestions for how the current value-added system might be made fairer and more rigorous.

[1] Further information about the Headteachers who were interviewed is given in the Reference section.
[2] For a discussion of the terms "error" and "bias" as used in this report see Appendix 7.

The Sizes of Primary Schools

About a third of all primary schools in England have a Year 6 cohort of less than 20 pupils. A cohort of 60 is unusually large and only about 15% have a Y6 cohort so large, whereas for a secondary school it would be considered small. Cohort sizes can also vary enormously from year to year in the same school, as one head teacher of a rural village primary school notes:

"Our present Year 6 only has 19 in it, but Year 5 running on behind them has 31. It's very rare that we get a cohort that's above 30." (H1)

It is questionable how much can be deduced from the results of a very small number of pupils:

"It does feel a little bit like a straw poll at times. It depends upon so many factors, on the day of the SATs, and that cohort of children, and it can make a huge difference, and when so much emphasis is put on these high stakes results it really is quite worrying." (H1)

"Because there are just 18 of them, statistically that can be tricky." (H6)

In Appendix 4, we show how modest differences in Key Stage 1 maths marks (of the sort that might be expected if one test were replaced by another similar one but with a different set of questions) can make a significant difference to a school's value-added measure at Key Stage 2. These differences are bigger for schools with smaller cohorts, as would be expected, because a single pupil makes a proportionately bigger contribution to the average score of a smaller school. If we were also to take into account likely uncertainties in reading and writing marks at KS1, as well as maths, English and science marks at KS2, the effects would be even larger.

The DfES did not publish value-added measures for schools that had 10 or fewer pupils eligible for KS2 assessment at the time of the 2003 tests. Our results lead us to suggest a more realistic minimum cohort size, to give reliable measures of schools, would be 50 (see Appendix 6). In other words, most primary schools are too small for their KS2 value-added measure to be secure enough to publish. We expect many primary schools' value-added measures to fluctuate wildly from year to year, even with consistent teaching.

Another important way in which primary and secondary schools differ is that primary-aged children are often taught across the curriculum by a single teacher for a whole year or longer. There is a danger that publishing a primary school's value-added measure holds up for public scrutiny the effectiveness of 3 or 4 teachers. In practice, the results at the end of Year 6 may be seen as being the responsibility of an individual teacher and individual teacher performance should be a private matter for the teacher, school and LEA to monitor and review.

The Quality of the Key Stage 1 Data

There is much concern among head teachers regarding the quality of Key Stage 1 data, and in particular whether KS1 results from infant schools and all through primary schools are comparable:

"There can be an issue, because where you have separate infant and junior schools, the infant schools without doubt push their children and their SATs results for Key Stage 1 further, whereas primary schools with all through Key Stage 1 and Key Stage 2 don't. There are some heads who would deny that, but I know from speaking with colleagues that it is actually true. My local inspectorate has said that it is recognised that this is an issue where you have a junior school next door to an infant school … and the progress appears to be lower than you'd expect." (H2)

"They [class teachers] feel up to a point they're on a hiding to nothing, which is compounded by the fact that we have a separate infant school next door, so they push the results anyway. We get children who come to us, they join us in year 3, we are actually left scratching our heads, thinking how on earth did that child get that level. We do our own assessments when they join us, we use NFER tests, and you get children who are getting perfectly good SATs levels who are getting very low standardised reading scores and standardised maths scores." (H2)

"The difficulty of getting results from other schools is that you have to accept that they're true. I have a girl in year 5 who is from a traveller family, who is on the special needs register and has some difficulties. I've got records for her … that apparently at Key Stage 1 SATs she got a level 2 across the board. Now, her value-added will not be good, because I have to accept that when she was 7 that was the level she was at." (H3)

"I have actually got some pupils where I have cognitive ability data and other standardised data which tells me that they are on standardised scores in the late 70s on a norm of 100, yet somehow those pupils have achieved a level 2A on the day, and therefore statistically have to achieve level 5 at the end of Key Stage 2, even though they are clearly well below average in ability." (H4)

"I think there's also an issue where you've got through primary schools and separate infant schools …We understand, because we're in a primary school, what it means to be a real level 3 or a real level 4 at Key Stage 2, and there's a lot more to it than is sometimes seen in a school that only has Foundation Stage and Key Stage 1. And also they pass the children on, so they don't have to worry about value added. And of course Key Stage 2 are marked externally, whereas Key Stage 1, although they are moderated, the papers are marked in schools … so I do have concerns about that." (H7)

"The children that are in my current year 6, when they were in year 2, we had a lot of children who got a level 4 for reading, and a lot of children who came in from this other infant school, a lot of them had level 4 for reading, and one or two for writing and maths as well, and yet there's no way that they were able to do level 4 work, but of course they will affect our value-added quite heavily, because if they got a level 4,

the most they can get is a level 5, so it's going to look as if they've only moved one level." (H7)

Some of the above concerns would be allayed, at least partially, by having KS1 tests marked externally. Clearly, external marking is considered important enough for the KS2 tests and it is quite proper to have concerns about comparability when internal marking is involved. But is there any hard evidence to suggest that the standards applied to the KS1 results vary across schools? Very little; because there has been no published investigation. However, it is interesting to note that Appendix 6 generates some unexpected relationships in the data, which might be explained by variations in standards applied to the KS1 assessments.

It might be helpful if the DfES were to investigate whether infant school KS1 results tend to be better than those of comparable all through primary schools, and to publish their conclusions. If no statistically significant difference were to be found, this might also allay many teachers' concerns.

There is official recognition that changes may be desirable to the way KS1 achievement is assessed. The DfES is trialling a new approach to assessing seven-year-olds in 2004 (DfES 2003c). However, the proposed changes will not enhance the KS1 data so far as value-added is concerned. Indeed, it seems as though more variation across schools might be introduced. This does not augur well.

Even assuming relatively reliable KS1 and KS2 results, value-added figures are always going to be less reliable than either of the individual results from the end of the key stages because they are based on two scores. The errors in each of these will necessarily combine to give a bigger error in the resulting value-added scores.

As noted earlier, we have investigated the effect on schools' value-added measures of introducing small differences into Key Stage 1 maths test marks. We find that the effects can be considerable and that one school can be affected much more than another school of similar size.

In summary, if value-added measures at KS2 are going to be trusted, issues with the quality of the end of Key Stage 1 data and teachers' concerns about the comparability of KS1 results from different kinds of schools need to be addressed. Until or unless consistent reliable data are available at the end of KS1, value added-scores at the end of KS2 are going to be problematic.

The Issue of Turnover

Value-added scores are published on a per-school basis. Because of that, it is natural for an uninformed observer to assume that the score for a school measures the efforts and skills of the staff within that particular school. But the score is actually a measure of the progress made from KS1 to KS2 by the pupils who happened to be attending that school at the time of the KS2 tests, and may have arrived there from other schools at any point during the previous 4 years. Families with primary school children are amongst those most likely to move house, and so the proportion of pupils who have been in the same school for all of Years 3 to 6 may not be very high:

"By the end of Year 6, only 60% of our pupils are what I'd call indigenous … some come in quite late … as late as Year 6 itself or Year 5, so we've only really had quite a short impact upon them." (H1)

"Even in a school like this, which is in a fairly stable suburban area, we only have 70% of the pupils who take the test at the end of Key Stage 2 who have been with us all the way from Key Stage 1." (H4)

Some schools have a particularly high turnover, for example, because of parents working in the armed forces:

"I have a school with high mobility, because of service children and other children that move into the area … Between 30 and 40% of my children are service children, sometimes it's been up to as high as 50% - it just depends on army movements … you just get children arriving - children arrive and leave throughout the year here … all the time there are children coming and going." (H3)

"It's demoralising [for class teachers], they work very hard, they already have to deal with the fact that they get children in who have difficulty - are already on the special needs register - and they work really hard with them, and feel they're just beginning to make progress, and then they move away." (H3)

[Referring to forthcoming Ofsted inspection]"I know that once again I will spend a lot of time talking and explaining and justifying and challenging things that are said to me … you do have to spend a lot of time actually making sure that they do understand the implications of that [high turnover]." (H3)

Some other schools have a high turnover because a significant proportion of parents are involved in higher education or academic research, and this can also mean a significant number of children come temporarily from abroad:

"It [turnover] tends to be more academic families where parents are at the University or training for the Church of England … we also get quite a large number of children who come who don't speak any English, and actually they make tremendous progress …some of them affect our Key Stage 1 results, that lowers our Key Stage results to some extent, they come out as Working Towards, because if they've just arrived from abroad there's no way they're going to get a level 1 or a level 2, but then they don't count when you get to the end of Key Stage 2, because they do the value-added on individual children, not on overall percentages, and those

children are not in our Year 6, because by then they've gone back to China or Korea or wherever they've come from. But in actual fact they make unbelievable progress." (H7)

The DfES does publish a *stability indicator* (see Appendix 1) along with the value-added measure for each child in an all through primary school (where children can move from the infant department to the junior department within the same school), but unfortunately not for junior schools.

A better stability indicator might be one that measured the proportion of pupils at the end of Year 6 who were in the same school at the start of Year 3 (rather than at the end of Year 2), as it could then be applied to junior schools as well. Even so, a stability indicator cannot tell the whole picture, because (as in the examples from head teachers above) the children who arrive and leave are often not a typical subset of all the pupils in the school, either academically or socio-economically.

As it stands, value-added measures are published if data are available for half the pupils in Year 6 – including those who have arrived from other schools. In other words, the published value-added measures can be based on less than half the pupils that the school taught in Years 3 to 6. We also note that for many schools, a great deal of effort, perhaps disproportionate effort, is put into working with transient pupils.

We suggest that, in future, the DfES uses a stability indicator that can be applied to junior schools, and only publishes value-added results for schools whose stability indicator is 90% or more.

Methodology

It would be easy to get lost in the mathematics and statistics of how value-added measures are calculated, but the devil is in the detail and the methodology chosen really matters in terms of how the performance of schools is judged. In this section we have broken down the key issues into digestible chunks, and have separated out much of the technical detail into appendices to the report. As one head notes:

"As far as the parents are concerned, I don't think I could begin to explain the maths to them, so I haven't tried." (H2)

Some of the head teachers we have spoken to have looked at the Government's methodology and the ways in which the results may be used, and have serious concerns:

"I was left in no doubt that statistically and mathematically it just doesn't stand up." (H2)

"Such an analysis could hide a less than effective Key Stage 1 - a very good value-added might not necessarily be a good thing" (H8)

Overview

The DfES calculates value-added measures using the methodology described in Appendix 1. To summarise:

The Key Stage levels achieved by a pupil at KS1 in maths, reading and writing are each converted to a points scale, and these three points are averaged to yield the pupil's Key Stage 1 average point score (KS1 APS).

A similar calculation is performed using maths, English and science achievements at Key Stage 2, giving the pupil's Key Stage 2 average point score (KS2 APS).

The KS1 APS results are divided into 15 bands, from lowest to highest. For each band, the *median* KS2 APS is calculated for all pupils nationally in that band. The median score is the middle score, in the sense that, if all pupils were lined up in order of score, it is the score below which there are half the pupils.

Each pupil is assigned an individual value-added score calculated as:

value-added score = (KS2 APS) – (median KS2 APS for pupils with same KS1 APS)

This means that a pupil who did better at KS2 than more than half of pupils with similar KS1 results scores positive, whereas one who did worse scores negative. The value-added score will be more positive (or negative) for pupils who did particularly well (or badly) at KS2 compared with the median.

A school's overall value-added measure is then calculated by taking the average (arithmetic mean) of the value-added scores of all the pupils in the school at KS2. So

if one child has progressed more than expected, and another has done less well, these will counterbalance each other to some extent. Most schools, as a result, have a measure in the range from roughly –2 to +2. The DfES adds 100 to all these scores before publishing, so the range becomes roughly 98 to 102.

Converting Key Stage Levels to Points

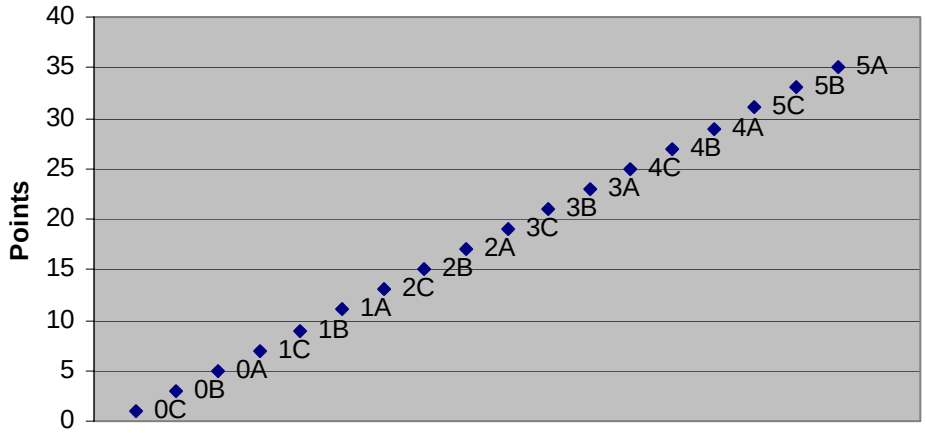
Our first observation is that the conversion from Key Stage levels to points should be such that the progress of a less able pupil can be compared with the progress of a more able one in a reasonably fair way.

Suppose two pupils both do better than expected at KS2 (compared against the national median line), but one is higher achieving than the other: how do we decide whose value-added score is bigger? Is getting a level 3 when level 2 is expected the same as getting a level 4 when level 3 is expected? The KS2 point scale (Table A1.2) has linear 6-point steps between each KS2 level, so the DfES's value-added calculation would give these two imaginary pupils the same score, which seems reasonable.

Point scores at KS1 are perhaps less of an issue for value added, as a pupil is only compared against others with the same KS1 score. However, as scores are averaged across 3 subjects, the average is probably more meaningful if the KS1 level-to-points scale is also linear. Table A1.1 shows that, in the DfES's current scheme, this is the case.

In fact, although it's not an essential requirement, the point scales used for KS1 and KS2 are essentially the same. If we were to split each level into A, B and C, at both Key Stages, we could plot points on a nice simple straight line graph, and these points would apply equally to both Key Stages. See Figure 1.

Figure 1 – KS1 and KS2 Levels and Points



However, there is a potential problem, because the A, B and C sub-levels are only used for level 2 at Key Stage 1. At all other KS1 levels, and for all KS2 levels, the whole level is lumped together as if it were a B. This distorts the calculation by making some pupils appear to have achieved extra progress and some less, as one head teacher observes:

"If, for example, you have children who've come up from a level 2C, and they get level 4, they automatically get counted as a level 4B, so anybody who comes up with a 2C has automatically made 14 points progress: they may not actually have got a 4B, they may have been worth a 4C, but there's no way of knowing because there are no A, B and C grades at Key Stage 2." (H2)

Whilst this effect might average out for a very large school, it will often distort the results of a school with small cohorts.

Validity of Comparing Key Stage Levels at Different Ages

There is an implicit assumption above, that a Key Stage level means something consistent, whether observed at Year 2, 3, 4, 5 or 6, but many head teachers have serious misgivings about this:

"People seem to believe that the national curriculum was designed in 10 levels and that they stepped nice and evenly from one level to another … this is a nonsense, so what you end up with is a level 3 at Year 3 being extraordinarily harder to achieve than a level 3 at Key Stage 1." (H4)

"Any primary school teacher, particularly at Key Stage 2, will tell you that any child who gets a level 3 on a Key Stage 1 test is most unlikely to get a level 3 on a Key Stage 2 test if they sat it at the same time … there is a mismatch between what is a level 3 at Key Stage 1 and what's a level 3 at Key Stage 2, four years later." (H2)

We can allay these head teachers' concerns to some extent. Although the points they make are perfectly valid, they don't affect value-added measures as long as the levels are consistent within each Key Stage. However, it does take us back to our points about the quality of Key Stage 1 data, and the possibility that teachers in all through primary schools may mark KS1 differently from teachers in infant schools, perhaps because of greater or lesser familiarity with the requirements of Key Stage levels at KS2.

The Ceiling Effect for Most Able Pupils

One of our most serious concerns with the DfES's methodology is that it exhibits a prominent ceiling effect, which adversely affects the most highly achieving children. This results partly from the fact that sub-levels are not used at KS2, but also because level 5 is the highest that can currently be achieved at KS2, and level 5 is also the median level nationally at KS2 for pupils who achieved level 2A or level 3 at KS1 (Barber 2003). This is made clear by some head teachers:

"If you have children coming up from infants already on a level 3, the maximum [increase in] … points they can get is 12, going from a level 3 to a level 5, because, for the purposes of value added, a level 3 is counted as a 3B and a level 5 is counted as a 5B, so whatever the children do, they can't make more than 12 points progress ... they couldn't add any value at all." (H2)

"For those 2A's and 3's children [KS1 results] it is impossible for you actually to give yourself a plus point in added value … Ironically this probably impinges in a small school more than it does here, … but nevertheless we are producing outstanding Key Stage 1 results, and a significant proportion of our cohort we are unable to add value to at all." (H4)

"The children who do show most value-added are our brightest kids, and they're the ones who are being scored nil, zero or satisfactory, and they really, by our own measures, and through PIPS, accelerate quite considerably, particularly as they approach Year 6. It seems ridiculous that the children who can indicate a positive value-added are not able to do so … it's not just a ceiling effect that stops children progressing and showing value in an ordinary way, it stops some very, very bright children from indicating so." (H8)

"It doesn't help schools where we get very high results at Key Stage 1 and Key Stage 2 - it's much harder to show value-added because we do very well early on … and as I say if you've got high results at Key Stage 1 it's hard, because now you can't even get a level 6 now on the SATs at Key Stage 2." (H7)

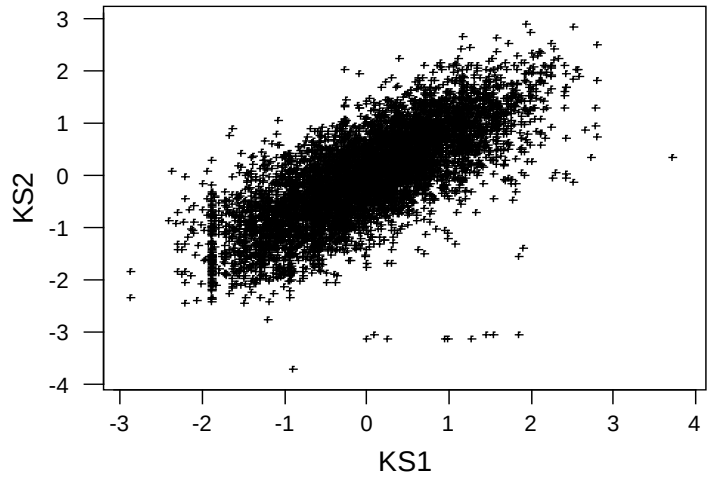
One might propose re-introducing level 6 at KS2, but some head teachers are concerned about issues around this:

[When there were tests for level 6] "There is no doubt that the test paper that we were given for our most able pupils was harder than the one that was put in front of the end of Key Stage 3 pupils that year. There seemed to be a determination that children weren't going to get level 6 at primary school. And now, of course, it's been abolished completely as an option for those more able pupils." (H4)

We must emphasise that this ceiling effect is not just an academic one: we have examples in our database of schools where half the pupils achieved 19 points or more (level 2A or higher) at Key Stage 1. (See Table A3.1 in Appendix 3.) It is mathematically impossible for these children to make a positive contribution to their school's value-added measure, however much progress they've made over 4 years. This simply cannot be fair.

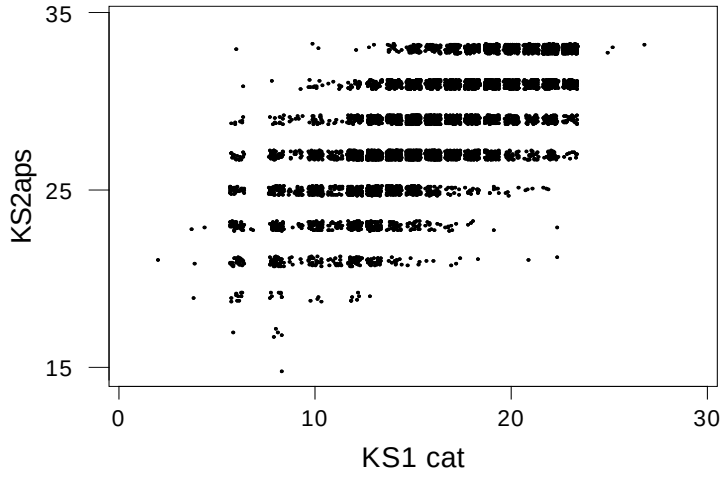
The ceiling effect is clearly seen when comparing the two scattergrams in Figures 2 and 3. Using marks, the relationship is shown in the scattergram below:

Figure 2 – Scattergram of KS2 marks against KS1 marks



The chart shows the familiar ellipse resulting from well-distributed scores at the ends of the two Key Stages.

Figure 3 - Scattergram of points derived from levels at KS2 against KS1



The scattergram in Figure 3 shows an ellipse which has had a ceiling put on it, which able pupils have hit at the end of KS2. The chart also shows a ceiling on the KS1 results and a floor on the KS1 data.

Using a Median Line

In the DfES's methodology, there is considerable use of arithmetic means (commonly referred to as averages), e.g. to combine an individual pupil's points from 3 subjects (adding the 3 point scores and then dividing by 3), or to obtain a school's overall value-added score from those of its pupils. But, when it comes to deciding what a pupil is expected to get at KS2, the DfES uses the *median* score for pupils with

similar KS1 average point score, instead of the arithmetic mean. Value-added measures are widely used throughout the world and have been very extensively studied over decades of research. To the best of our knowledge, no university statistician, educational researcher or any other educational system recommends or uses a median line when studying or calculating school value added scores, and we have failed to find any justification for its adoption.

This distinction may sound academic, but Critchlow and Coe (2003) and Critchlow (2004) have shown, through detailed analyses, that using the median biases the results. Those schools with mostly high-achieving children have their value-added scores pushed down, and those at the other extreme have their scores boosted. We won't duplicate the above authors' arguments here, but we note that our own analysis, using data we have at the CEM Centre on 6175 pupils in 349 schools who sat KS2 tests in 2003, entirely corroborates their findings. In Appendix 3, we show that using the median produces value-added measures for schools that are biased in favour of schools with lower KS1 results by up to 2 points, compared with using an arithmetic mean. If schools in an LEA are ranked in a league table of value-added measures, then using median instead of mean can make a big difference to some schools' placings: differences in ranking of more than 50 places out of 300 are quite possible. Many schools branded as below average may, in reality, have performed above average, and vice versa.

Despite these serious problems associated with using what the DfES calls the *national median line*, there is a high risk that teachers will now feel compelled to closely measure and judge their own pupils' progress against it in inappropriate ways, as one head teacher explains:

"The inevitable result of all this is that schools will need to do their pupil tracking in a different way to the way I've traditionally done it. What we're going to need to do is to do our pupil tracking against the median line trajectory from Key Stage 1 to Key Stage 2 ... what that means is that one's pupil-tracking is inevitably going to be done now against National Curriculum levels, as measured by the optional SATs. Now the optional SATs were never ever designed to have such high stakes attached to them as this process now attributes. So what you end up with is Year 3 teachers in a terrible state, to be frank, because they find that at the end of the year with them they've got, particularly at the lower end of ability, a lot of children who have made no progress at all, it would seem ... or who have even gone backwards, because those children, faced with the very pupil-unfriendly and little-trusted optional SATs, have found it incredibly difficult. And the problem that you've then got is that you've got teachers' performance management attached to 3 elements, and the first of them is pupil progress. So clearly what schools get railroaded into by this value-added measure is a kind of tracking against a median line throughout Key Stage 2." (H4)

Points or Marks

An alternative approach to the DfES's methodology would involve marks instead of levels.

To fairly compare marks in different subjects it would be necessary to standardise (i.e. re-scale) the marks in each test, for example, to have a mean of 50 and a standard

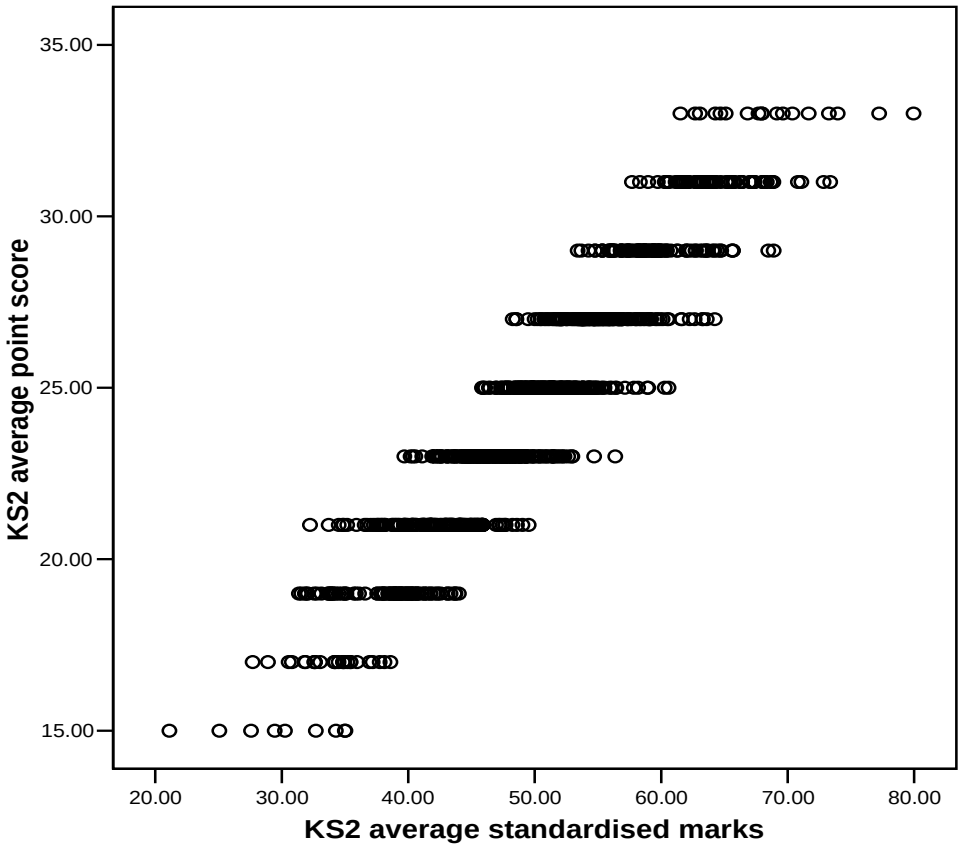
deviation of 15. It would then be possible to average marks across subjects, and to use a national regression line instead of the existing national median line. The scattergrams in Figures 2 and 3 suggest that the mark approach would have advantages over the point approach – even if a median line were used.

A further difficulty with using points was first pointed out to us by Critchlow (private communication). There are only 4 distinct point scores a pupil can be assigned in each of maths, English and science at KS2. (See Table A1.2 in Appendix 1.)

If one looks at how a pupil's average of these 3 point scores relates to his or her average standardised marks in the 3 subjects, a chart looking like Figure 4 below is the result. We have used the published Level Threshold Tables for KS2 in 2003 (QCA 2003), and computer-generated quasi-random marks for 1000 imaginary pupils. (The same effect is observed using real pupil marks.)

We can see from Figure 4 that there is an overlap between the range of standardised marks for adjacent point scores: two pupils having the same average standardised marks can have average point scores differing by 6 points. This leads us to doubt the appropriateness of grouping children into only 4 distinct point scores for each subject before averaging.

Figure 4 - KS2 average point score vs. KS2 average standardised marks



It is clear that children with relatively high average marks can get a low point score, and vice versa. But since the results of the value-added exercise do not report figures for individuals, this would not matter provided the apparent anomalies did not

influence the school score unduly. For this to happen there needs to be a sufficient number of children in the cohort, and this brings us back to the issue of size of cohort for which it is fair to publish results. The link between marks and points underscores the need to exercise caution.

Contextual Effects

Value-added measures are intended to compare like with like: by looking at similar pupils in different schools we can hope to compare their progress. But if schools are to be held to account for their performance, we must compare similar schools with similar schools. It would surely be unfair to hold schools to account for their progress in a deprived area by comparing their results with one in an affluent area. Nor would it seem right to hold to account for the progress of its children a school with a high proportion of children whose first language is not English, by comparing them with one where all the children have English as their first language.

There is evidence (Appendix 3) that the way the DfES currently calculates value-added may be biased against schools with a more able intake, as one head teacher observes:

"If the figures were corrected, the league tables would look very different … Schools working with children in more deprived circumstances, and schools like this, work just as hard and produce similar results, but I think the league tables have skewed that." (H2)

Another issue arises where special needs children are integrated into a mainstream primary school:

"The value added, where it comes out as some sort of ranking in league tables, doesn't actually indicate the range of children … The children might be included, and it's a better way than the threshold, but it doesn't give recognition to the work that's carried out with the special needs children … You simply can't compare the progress that those children make with progress of all schools or a mainstream cohort." (H5)

Furthermore, if any attempt is going to be made to track the value-added measure of individual schools over time, it will be necessary to take into account that a school's circumstances are not static:

"All of the terraced house in the City which used to traditionally have families are now, most of them, students. We've certainly noticed a big change in our intake which may have an effect on value-added long term." (H7)

Appendix 6 indicates that models used to calculate value-added scores for schools can be enhanced by including whole school measures. Very different impressions of some schools are obtained when the average intake to a school is considered in addition to individual pupils' results. If schools are to be judged on the basis of their value-added scores then they must be fair and that means comparing like with like: like pupils with like pupils **and** like schools with like schools.

Conclusions

Our investigations have shown that the uncertainties and biases associated with value-added league tables at Key Stage 2 are sufficiently great, and other issues such as turnover and school context so important, that it's difficult to regard the published 2003 value-added for primary schools as likely to have a positive impact. And, as one head teacher notes:

"There's also something about consistency of results as well, that often if you're fairly consistent throughout, the value-added doesn't need to be massively high – there's some argument that a sort of average value-added means that you're performing fairly consistently … so when you do get very high value-added there's some things you need to look at before you make too many conclusions." (H8)

There is a useful distinction to be made between the use of value-added for accountability purposes (school performance tables) and for school improvement purposes (professional use within the school). It is doubtful that many readers other than statisticians and some teachers are in a position to interpret the currently-published tables in any fair and meaningful way, and many schools are already involved in useful ways of measuring their pupils' progress.

If value added measures are going to be published in school performance tables along similar lines to the 2003 exercise, then we believe they should, in future:

1. Indicate the uncertainties in schools' scores much more clearly and prominently. The Royal Statistical Society (RSS 2003) has called for measures of uncertainty to be reported whenever performance data are published.

2. Address the issue of the quality of Key Stage 1 data.

3. Use Key Stage sub-levels (A, B and C) for all levels or standardised test marks.

4. Use a mean line or a regression line instead of a median line.

5. Not publish data for schools with cohort size less than 50 or stability indicator less than 90%.

6. Use a stability indicator that can be applied to all schools.

7. Give prominence to the different circumstances in which schools find themselves.

Taken together we are aware that this list suggests that value-added measures should not be published at the end of KS2 in 2004. We are also aware that, even if the issues were addressed, it would be several years before value-added measures could be published, and then only for some schools. We recommend that the whole issue of performance tables for primary schools should be re-examined.

In Appendix 7 we have included some individual head teachers' conclusions.

References

H1. Telephone interview with Mrs Sue Sayles, head teacher of Riccall Community Primary School in North Yorkshire, February 2004.

H2. Telephone interview with head teacher of a junior school in South London, February 2004.

H3. Telephone interview with head teacher of a primary school in the Home Counties, February 2004.

H4. Telephone interview with Mr David Pratt, head teacher of Little Common School, Bexhill, East Sussex, February 2004.

H5. Telephone interview with head teacher of a primary school in an economically deprived area in North West England, March 2004.

H6. Telephone interview with head teacher of a small primary school in North West England, February 2004.

H7. Telephone interview with head teacher of a junior school in North East England, February 2004.

H8. Telephone interview with Mr George Barber, head teacher of Yarm County Primary School, Stockton-on-Tees, February 2004.

Barber, G. (2003) *League tables that penalise perfection*, Letter in the Times Educational Supplement, October 24th 2003.

Critchlow, J. and Coe, R. (2003) *Serious flaws arising from the use of the median in calculating value-added measures for UK school performance tables,* International Association for Educational Assessment (IAEA) Annual Conference, October 2003. www.aqa.org.uk/support/iaea/papers/critchlow-coe.pdf.

Critchlow, J. (2004, submitted for publication) *Problems arising from the use of the median in the proposed KS2 value-added tables*

DfES (2003a) *How to read the tables*, www.dfes.gov.uk/performancetables/primary_03/p4.shtml.

DfES (2003b) *Value added technical information*, www.dfes.gov.uk/performancetables/primary_03/p5.shtml.

DfES (2003c) *Excellence and Enjoyment – A Strategy for Primary Schools*, www.dfes.gov.uk/primarydocument/.

Fitz-Gibbon, C. T. (1997) *The Value Added National Project: Final Report: Feasibility studies for a national system of Value Added indicators* (COM/97/844). London: School Curriculum and Assessment Authority.

Goldstein, H. (1987) *Multi-level models in Educational and Social Research.* London: Griffin School Effects.

Goldstein, H. (2003) *A commentary on the secondary school value added performance tables for 2002,* www.ioe.ac.uk/hgpersonal/value-added-commentary-jan03.htm.

Goldstein, H. (2004) *Value added – a commentary on the KS1-KS2, KS2-KS3 league tables December 2003 and KS3-KS4 January 2004*, www.ioe.ac.uk/hgpersonal/value_added_for_primary_schools_dec_2003.htm.

Harker, R. and Tymms, P.B. (in press) *The Effects of Student Composition on School Outcomes,* School Effectiveness and School Improvement.

QCA (2003) *Key Stage 2 level threshold tables 2003,* www.qca.org.uk/ages3-14/tests_tasks/2636.html.

RSS (2003) *Performance indicators: good, bad and ugly*, The Royal Statistical Society. www.rss.org.uk/archive/reports/231003.pdf.

Tymms, P. and Henderson, B. (1995) *The Value Added National Project Technical Report: Primary* (REF: COM/96/407). London: School Curriculum and Assessment Authority.

Appendix 1 – The DfES Calculation of Value-added at KS2 in 2003

The calculations performed to determine value-added scores for schools at KS2 in 2003 are described in DfES (2003b).

At KS1, each pupil is awarded a number of points from each of the reading, writing and mathematics tests, depending on the level they achieved.

In 1999, level 4 was the highest achievable, and so 27 was the highest possible point score from any test. In the case of reading, where there was a reading task and a reading comprehension test, the result from the comprehension test was used if the pupil entered the test and achieved level 3 or higher, otherwise the result from the reading task was used.

From these scores, the pupil's KS1 average point score (KS1 APS) is calculated as the arithmetic mean of the 3 individual point scores.

Table A1.1 – Levels to Points at Key Stage 1

KS1 Level	KS1 Points
Absent, Disapplied or Missing	Disregarded
Working Towards	3
1	9
2C	13
2B	15
2A	17
3	21
4+	27

At KS2, each pupil is awarded a number of points in each of English, mathematics and science, based on the level they achieved in each:

In 2003, level 5 was the highest achievable, so 33 was the highest possible point score in any subject.

From these scores, the pupil's KS2 average point score (KS2 APS) is calculated as the arithmetic mean of the 3 individual point scores.

Table A1.2 – Levels to Points at Key Stage 2

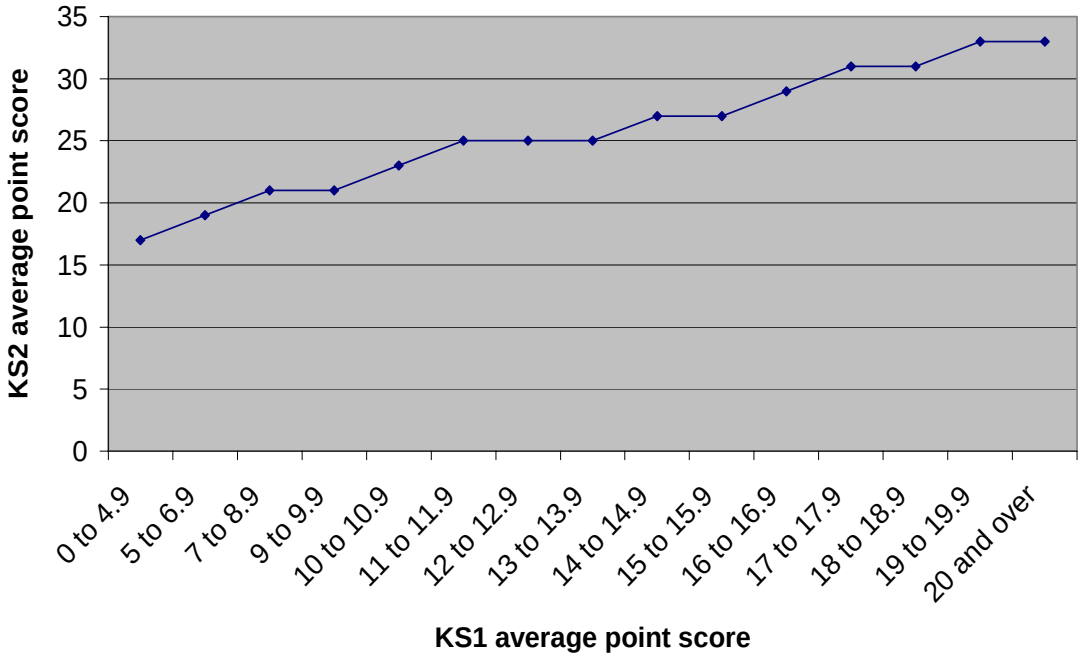
KS2 Level	KS2 Points
Missing, Lost, Ineligible, Disapplied, Absent or Annulled	Disregarded
Working below the level of the test	15
Not awarded a level	15
2	15
3	21
4	27
5	33

Each pupil's value-added score is calculated by comparing their KS2 APS with the median KS2 APS of all pupils nationally in 2003 who scored the same KS1 APS in 1999. The DfES uses the term *National Median Line* to describe the median KS2 APS scored by pupils with different KS1 APS, and publishes one National Median Line for mainstream schools and one for special schools. The mainstream line can be tabulated and drawn graphically as follows (DfES 2003b):

Table A1.3 – National Median Line tabulated for Mainstream Schools

KS1 Average Point Score	National Median KS2 Average Point Score
0 to 4.9	17
5 to 6.9	19
7 to 8.9	21
9 to 9.9	21
10 to 10.9	23
11 to 11.9	25
12 to 12.9	25
13 to 13.9	25
14 to 14.9	27
15 to 15.9	27
16 to 16.9	29
17 to 17.9	31
18 to 18.9	31
19 to 19.9	33
20 and over	33

Figure A1.1 – Graph of National Median Line for Mainstream Schools



Note that the *median* score of a set of pupils is the score below which half of the pupils appear, if all the pupils were lined up in order of score: it is not the same as the *arithmetic mean*, which would be calculated by adding together all the individual pupil scores and then dividing by the total number of pupils.

A pupil's individual value-added score is the difference between his or her KS2 APS and the median score for pupils who scored similarly at KS1. For example, a pupil with KS1 APS = 15 and KS2 APS = 29 would have a value-added score of +2, because the National Median KS2 APS for pupils with KS1 APS of 15 is 27.

A school's KS1 to KS2 value-added measure is calculated as the arithmetic mean of its pupils' individual value-added scores, and then adding 100. So, for example, with 5 pupils scoring +2, 0, -1, +1 and +2, a school would score 100.8. However, a school with less than 11 eligible pupils does not have its overall score published.

The DfES value-added tables include two additional measures of each school:

- A coverage indicator

  The percentage of pupils eligible for the KS2 tests who were actually included in the calculation. Those for whom no KS1 data is available, or who were absent for the KS2 tests, are excluded from the calculation.

- A stability indicator

  For schools with both an infant and a junior department, this shows the percentage of KS2 pupils who took their KS1 tests at the same school. This figure is not shown for schools that cater for the junior age range only.

Some guidance on comparing the value-added measures of different schools is contained in DfES (2003a), which says:

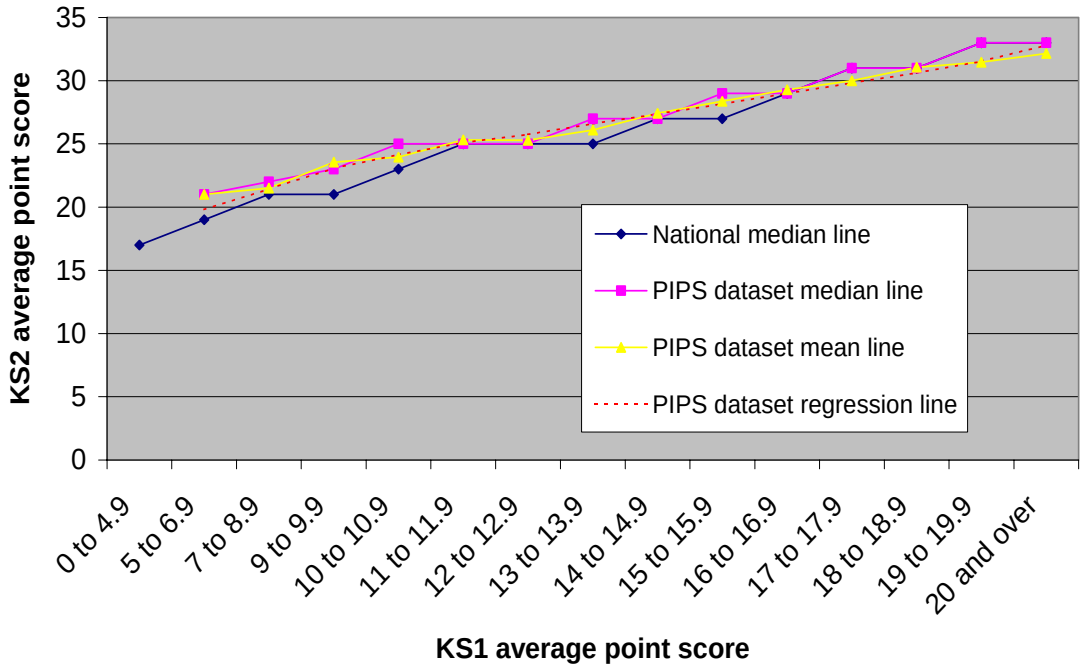
"When comparing schools with cohorts of about 30 pupils, differences of up to 1.3 should not be regarded as significant, while for schools with about 50 pupils, differences up to 1.0 should not be regarded as significant."

Corresponding differences for cohorts of 11 pupils are not mentioned in this DfES document, although measures for cohorts as small as this are published. See Appendix 4 for conclusions of ours regarding differences for small cohort sizes.

Appendix 2 – Medians, Mean and Ordinary Least Squares

Figure A2.1 below shows the KS1-KS2 2003 National Median Line, as published by the DfES, along with 3 other lines calculated from a sample of 6175 pupils from 349 primary or junior schools who were in Year 6 in 2003 and whose data were collected as part of the PIPS Project at the CEM Centre.

Figure A2.1 – KS1 to KS2 median and mean lines



The lines from the PIPS Project are:

- PIPS Dataset Median Line

  Calculated similarly to the DfES National Median Line, but using the dataset of 6175 pupils instead of all pupils in mainstream schools in England. The top 4 points on the Dataset Median Line are identical to those on the National Median Line.

- PIPS Dataset Mean Line

  Instead of calculating the *median* KS2 APS for pupils in the dataset with a particular range of KS1 APS, we have calculated their *arithmetic mean*.

- PIPS Dataset Regression Line

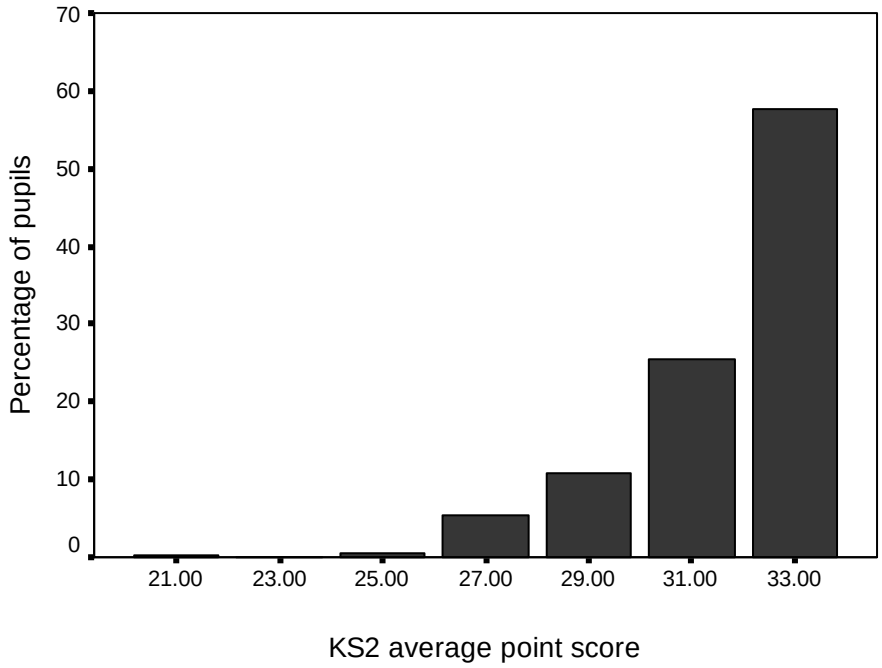
  Instead of using discrete ranges of KS1 APS, we have calculated the ordinary least squares linear regression line of KS2 APS against KS1 APS for all the pupils in the dataset.

Note that the Dataset Median Line is very similar to the National Median Line, suggesting that our sample of 6175 pupils is a reasonably representative one. Note

also that the Dataset Mean Line and the Dataset Regression Line are very similar to each other, and much smoother (less stepped) than either of the Median Lines.

It is instructive to look at the distribution of KS2 average point scores achieved by pupils with a given range of KS1 average point scores. Figure A2.2 below shows an example, using the 1120 out of 6175 pupils in our PIPS dataset whose average point score at KS1 was 19 or more.

Figure A2.2 – KS2 APS of Pupils with KS1 APS ≥ 19 in PIPS Dataset



This is obviously a very skewed distribution, but that's not surprising: we would expect most of the pupils who did well at KS1 to also do well at KS2, and the KS2 point scale is capped at 33 points. Because more than half of these 1120 pupils (57.7% of them) scored 33 points at KS2, 33 is their median KS2 APS. But their arithmetic mean KS2 APS is less than this: approximately 31.7 points. So, although none of these pupils scored more than the median at KS2, more than half scored more than the arithmetic mean.

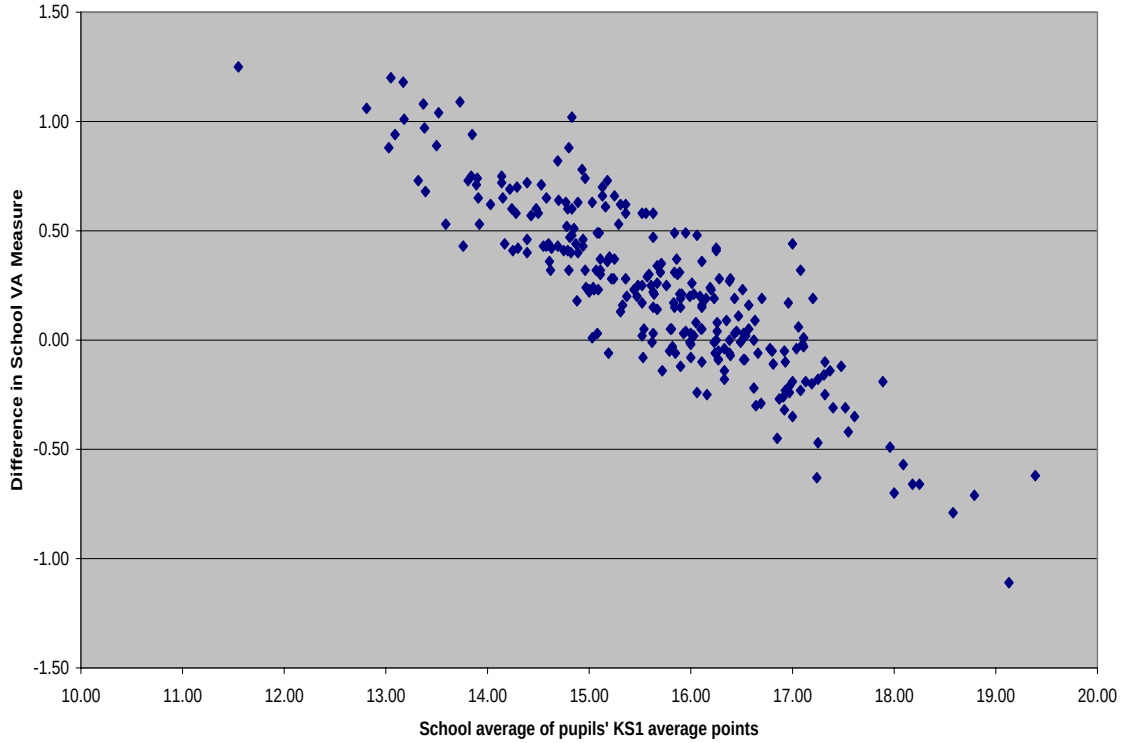
Appendix 3 – Value-added using Median or Mean Lines

The DfES calculates school value-added measures at KS2 using a national median line (Appendix 1), but alternative lines, which may be equally or more appropriate, could be used instead (Appendix 2). It is therefore useful to investigate how using different lines affects the value-added outcomes for different schools.

Previous studies by Critchlow and Coe (2003) for KS2-KS3 and KS3-KS4, and by Critchlow (2004) for KS1-KS2 have already highlighted problems which can arise from using a median line. They point out that using a median line doesn't produce self-consistent results: for instance, at KS4, the whole cohort's overall value-added score is well below zero, which doesn't make sense. They also identify that because the median is constrained to only a small number of possible values, small changes in the distribution of scores can cause large changes in the median.

From our database of 6175 pupils in 349 schools, we have calculated school value-added measures using (a) the DfES's national median line, and (b) our dataset's mean line. Figure A3.1 below shows the difference between these measures (i.e. a – b) for each school, plotted against the school's average KS1 APS. In the same way that the DfES excludes schools with cohort sizes of 10 or less from its published value-added tables, we have only shown the results below for the 265 out of 349 schools in our database for which we have matching KS1 and KS2 data for more than 10 pupils.

Figure A3.1 – Value-added difference, national median line vs. dataset mean line



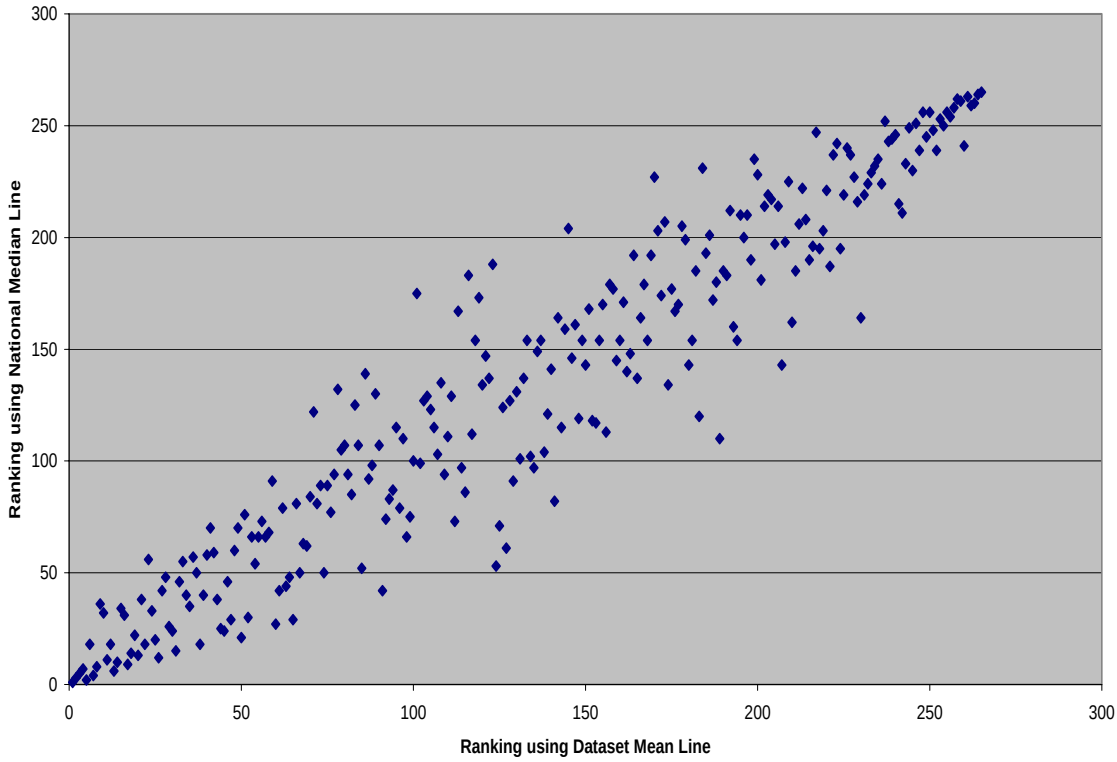
This scattergram shows a statistically significant, and visually very obvious, trend: compared with using the dataset mean line, the national median line gives value-added measures that are weighted heavily against schools with high average pupil attainment at Key Stage 1. Although this does not tell us which of the lines is the

better one to use, Critchlow (2004) has previously observed a very similar result using data for over 6,000 pupils in a single LEA, and has argued that the median line misrepresents the progress made by schools whereas use of the mean line would avoid this problem.

Referring back to Figure A2.2 in Appendix 2, we can see that, in our PIPS dataset, if we try to use the median KS2 APS as their "average" KS2 APS for pupils with KS1 APS of 19 or more, we come to the bizarre conclusion that none of these pupils did better than average. On the other hand, if we use the arithmetic mean instead of the median as the "average", we find that 57.7% of them achieved better than average, which seems more sensible.

To further illustrate the discrepancy between using median and mean lines in Figure A3.2 below, we have plotted the value-added rankings of the 265 schools from Figure A3.1, using the two different lines.

Figure A3.2 – School KS2 value-added rankings – median line vs. mean line



In this figure, the school scoring highest has position 1 (to the left or bottom), the school scoring worst has position 265 (to the right or top). What is striking is that many schools' positions in this league table differ by more than 50 places depending on whether the median or mean line is used. We would expect similar effects to be observed in individual LEA league tables, so that some schools whose results have been published as well below average in their LEA may actually be well above average, and vice versa.

To give a concrete example of the difference for a school of using a median or a mean line, Table A3.1 below shows KS1 and KS2 APS results for the 12 pupils in a school from our PIPS dataset.

We have calculated each pupil's value-added score in two different ways, using the DfES's national median line and the PIPS dataset's mean line.

Table A3.1 – KS1 to KS2 Value-added for Pupils in One School

	KS1 APS in 1999	KS2 APS in 2003	Value-added score (National median line)	Value-added score (Dataset mean line)
Pupil 1	21.00	33	0	+0.84
Pupil 2	19.67	31	-2	-0.46
Pupil 3	21.00	33	0	+0.84
Pupil 4	21.00	33	0	+0.84
Pupil 5	17.67	31	0	+1.00
Pupil 6	15.00	29	+2	+0.63
Pupil 7	21.00	33	0	+0.84
Pupil 8	21.00	33	0	+0.84
Pupil 9	21.00	31	-2	-1.16
Pupil 10	18.33	29	-2	-2.07
Pupil 11	19.67	31	-2	-0.46
Pupil 12	16.33	31	+2	+1.71

It is worth noting that:

- Every child in the cohort achieved better than the DfES "expected level" at KS2 (27 points, i.e. Key Stage level 4).
- Only 2 out of 12 children achieved a positive value-added score using the DfES's national median line, whereas 8 out of 12 scored positive using our dataset mean line.
- Half of the pupils were affected by the ceiling effect in the DfES's methodology, where those with KS1 APS ≥ 19 cannot be given a positive value-added score.
- This school was ranked in the top 20% within its LEA, based on aggregate KS2 results, but in the bottom 30%, based on value-added from KS1 to KS2.
- Within the PIPS dataset's 265 schools with cohort size over 10, this school's value-added ranking would rise from 183rd out of 265 using the national median line to 116th out of 265 using our dataset mean line.

This school (and others like it) is branded as worse than average in value-added terms, using the DfES's calculations. We seriously doubt whether this is a fair verdict.

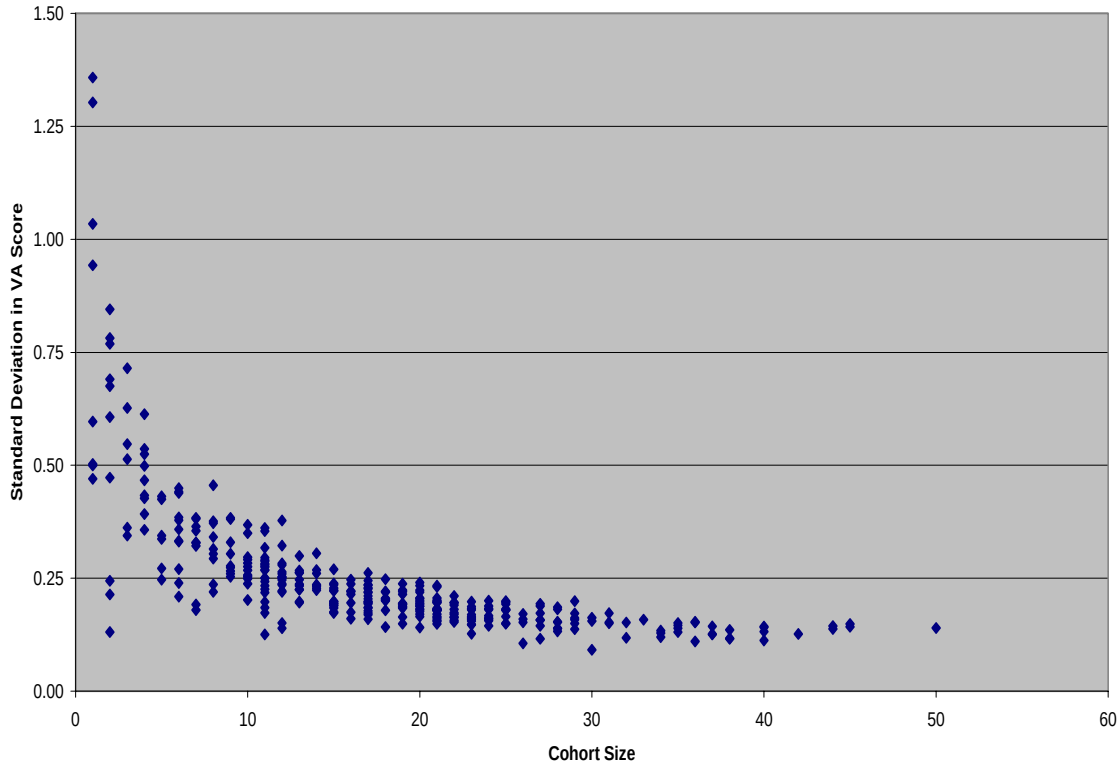
Appendix 4 – Effect on KS2 Value-added of Errors in KS1 Marks

No two tests designed to measure the same thing will produce identical results, but reliable tests should have a high degree of correlation. To illustrate the effect on value-added results of varying individual test results, we have investigated adding quasi-random errors into the marks scored by 6175 pupils in the maths test at Key Stage 1 in 1999.

We find that introducing normally-distributed errors with a standard deviation of 3.5 marks gives test results with a correlation of 0.9 to the original results: in the absence of a reliability[3] figure from QCA we have assumed that a figure of 0.9 is about right. By re-calculating each of 349 schools' value-added measure 10,000 times using different random errors distributed as described, we can obtain a good estimate for the resulting standard deviation in each school's value-added measure.

Figure A4.1 below illustrates our results, using the same National Median Line based value-added calculation as used by the DfES.

Figure A4.1 – School value-added standard deviations due to varying KS1 maths marks, using national median line



The standard deviations are largest for schools with smallest cohort sizes, as would be expected, because changes in the marks of one pupil can have a much bigger effect on the school's overall measure for small cohorts. Statistically, one would expect the standard deviation to vary roughly as $1 / \sqrt{}$ (cohort size), and the above figure roughly bears this out.

[3] For a discussion of the term reliability as used in this report see Appendix 7.

What is striking, however, is that the effects of varying test marks on one school with a given cohort size can be twice as large or more as the effect on another school with the same cohort size, especially for small cohort size. We attribute this to the fact that some schools will have more pupils whose attainment places them near to Key Stage level boundaries, and whose point scores are therefore very susceptible to small changes in marks scored in tests, whereas other schools will have more pupils placed nearer to the middles of different levels, where their level will be less affected by small mark changes.

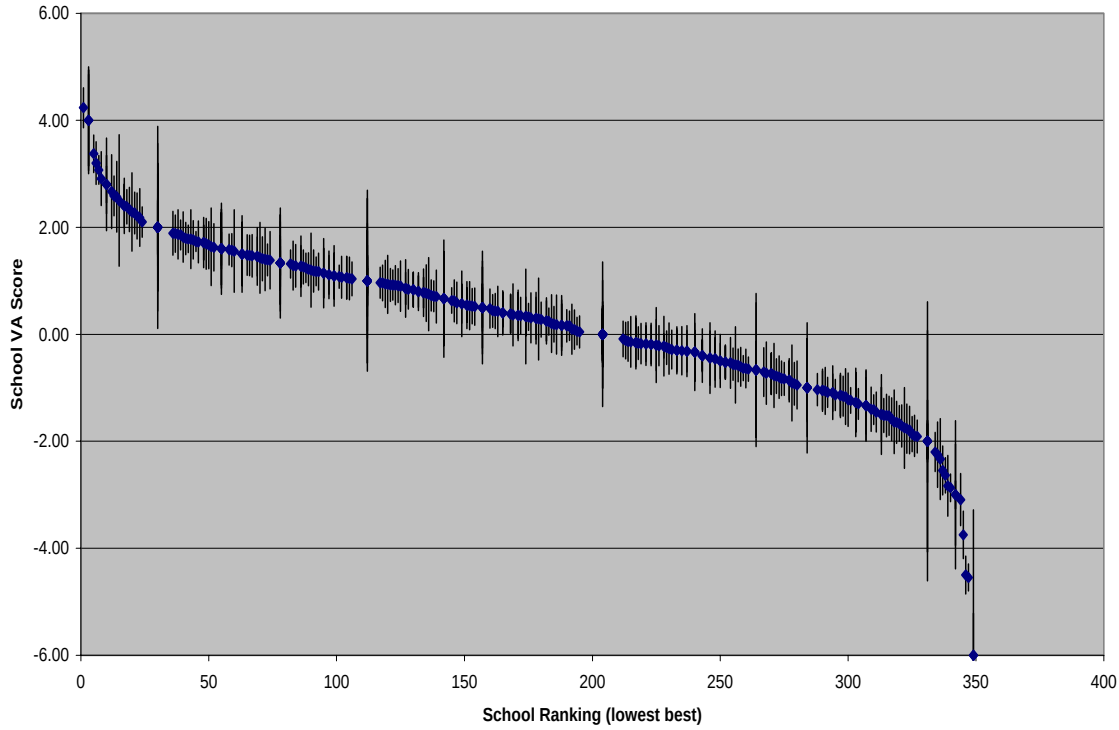
If we were to take into account the effect of errors in reading and writing scores at KS1, and in English, maths and science results at KS2, we would probably agree with the DfES statement (see Appendix 1) *"with cohorts of about 30 pupils, differences of up to 1.3 should not be regarded as significant, while for schools with about 50 pupils, differences up to 1.0 should not be regarded as significant"*.

The DfES publishes results for cohorts as small as 11, but doesn't say what the error bars are for these. Our results suggest that they are probably at least twice as large as for cohorts of size 30, making the value-added results of these small cohorts so unreliable that they really should not be published. Goldstein (2004) has made some important observations about the inadequate attention to uncertainty intervals in the results published by the Government.

Figure A4.2 below shows another way of looking at the results of introducing errors into maths KS1 marks. Here we have ranked our 349 schools in order of value-added score, and shown for each error bars of 2 standard deviations above and below (corresponding to a 95% confidence interval). We can see from this that schools apparently 100 places apart in this league table are in many cases indistinguishable within the limits of the errors. Similar uncertainties must be expected in league tables of value-added scores for individual LEAs, rendering positions within the tables largely meaningless.

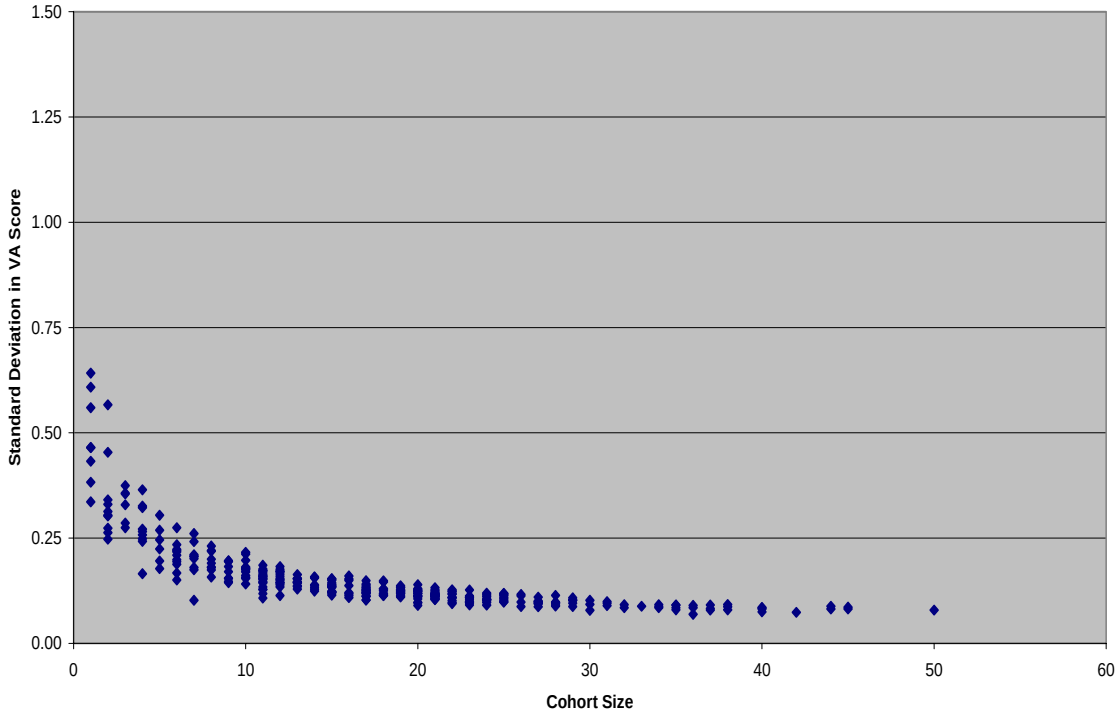
Figure A4.2 – Value-added scores (national median line) with KS1 maths error bars.

NB this assumes neither error in the other KS1 marks nor any error in the KS2 results



It is very interesting to look at the effect of random errors in KS1 maths marks again, but this time using value-added based on our Dataset Mean Line as shown in the Figure A4.3 below, instead of the National Median Line.

Figure A4.3 – School value-added standard deviations due to varying KS1 maths mark using dataset mean line

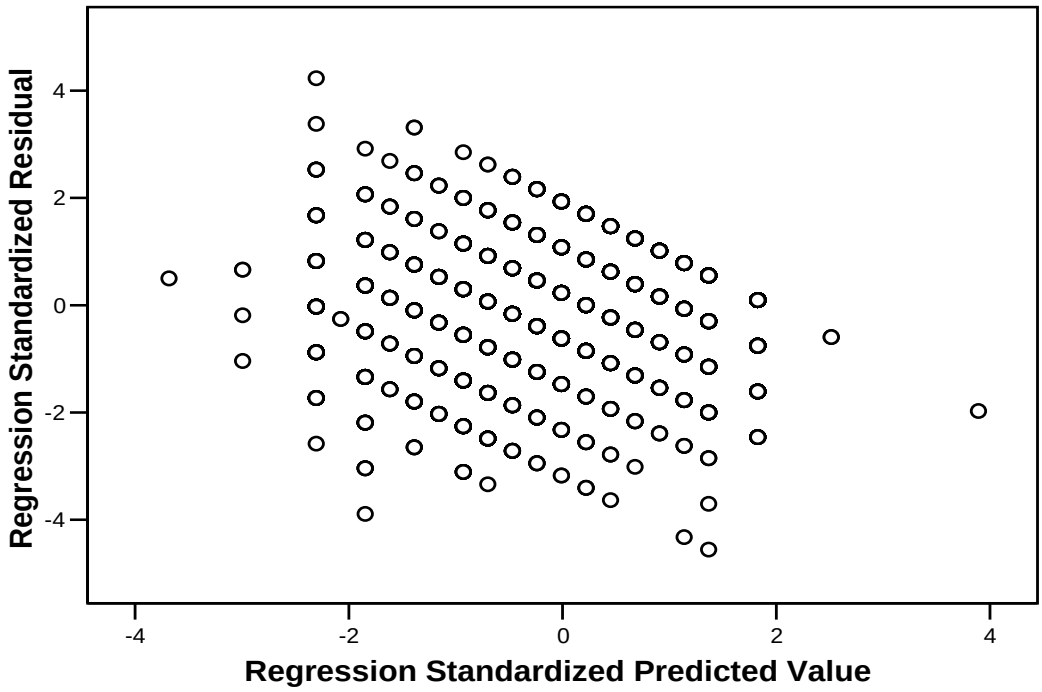


What is very noticeable is that the standard deviations here are significantly smaller. In other words, schools' value-added measures are less susceptible to errors in the individual test marks when using a Mean Line instead of a Median Line to calculate expected KS2 APS for individual pupils. This is presumably because the Mean Line is smoother (has less pronounced steps) than the Median Line. (See Appendix 2.) The steps mean that a small change in a pupil's KS1 score can cause a big change in that pupil's expected score at KS2. Even in a school with a large cohort, a proportion of the pupils will be close enough to the step-edges that errors in their marks will lead to a measurably bigger uncertainty in the school's overall value-added measure, compared with using a smoother Mean Line. This lends weight to the argument that using a Mean Line would be much more appropriate.

Appendix 5 – Ordinary Least Squares Regression Analyses

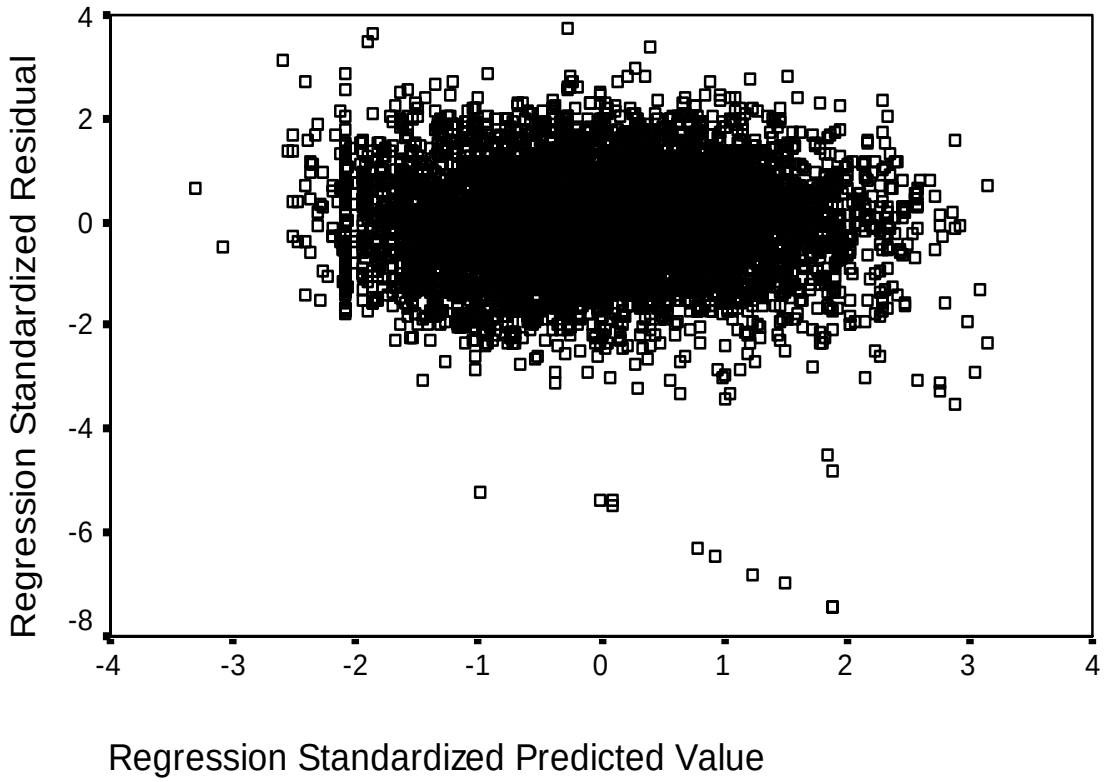
We have carried out a pupil-level linear Ordinary Least Squares (OLS) regression analysis, using the KS1 APS and KS2 APS scores of each of 6175 pupils in 349 schools. The correlation is good (R = 0.707), but when we look at standardised residuals, as shown in Figure A5.1 below, we can see that there is a lack of points in the top right quadrant. This corresponds to the ceiling effect discussed elsewhere in this report: pupils achieving Level 2A or Level 3 at KS1 are not permitted by the KS2 tests to achieve above level 5.

Figure A5.1 – Residuals scattergram for KS2 APS vs. KS1 APS



By way of comparison, we carried out an OLS regression analysis of the same pupils' mean standardised marks at KS2 against standardised maths, reading and writing marks at KS1 (treating the 3 marks at KS1 as separate variables). In this case we find a slightly better correlation (R = 0.753), and a residual scattergram which is much more symmetrical about the origin, as shown in Figure A5.2 below.

Figure A5.2 – Residuals scattergram for KS2 average mark vs. KS1 marks



Following on from the pupil level OLS regression analyses, we have also looked at the school level residuals and the standard errors in these. Figures A5.3 and A5.4 display the schools ranked by averaging the per-pupil residual points or marks over each school, and the error bars shown correspond to 95% confidence intervals for these average (i.e. mean) residuals.

Pupil marks here are based on standardising the marks for each KS1 and KS2 test to have a mean of zero and standard deviation of 1 mark, so the vertical scale in Figure A5.4 (standardised marks) is not directly comparable to that in Figure A5.3 (points). However, we have drawn the two figures with vertical scales so that the two ranking curves have approximately the same scale on paper: note that the error bars are smaller compared with the overall range of residual scores when using marks instead of points. In the points case, the error bar sizes are comparable to those quoted by the DfES (1.3 points for cohort of 30, 1.0 for cohort of 50, see Appendix 1), but are, of course, larger still for small cohort sizes (for which the DfES does not quote error sizes).

Figure A5.3 – School Ranking using Points with 95% confidence intervals

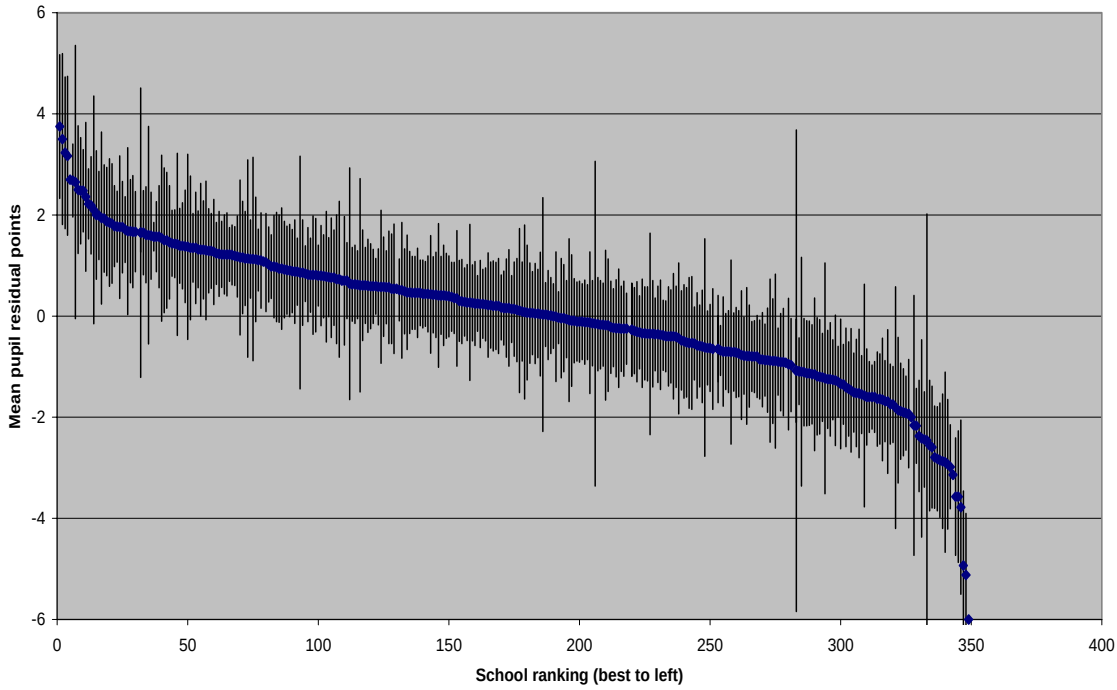
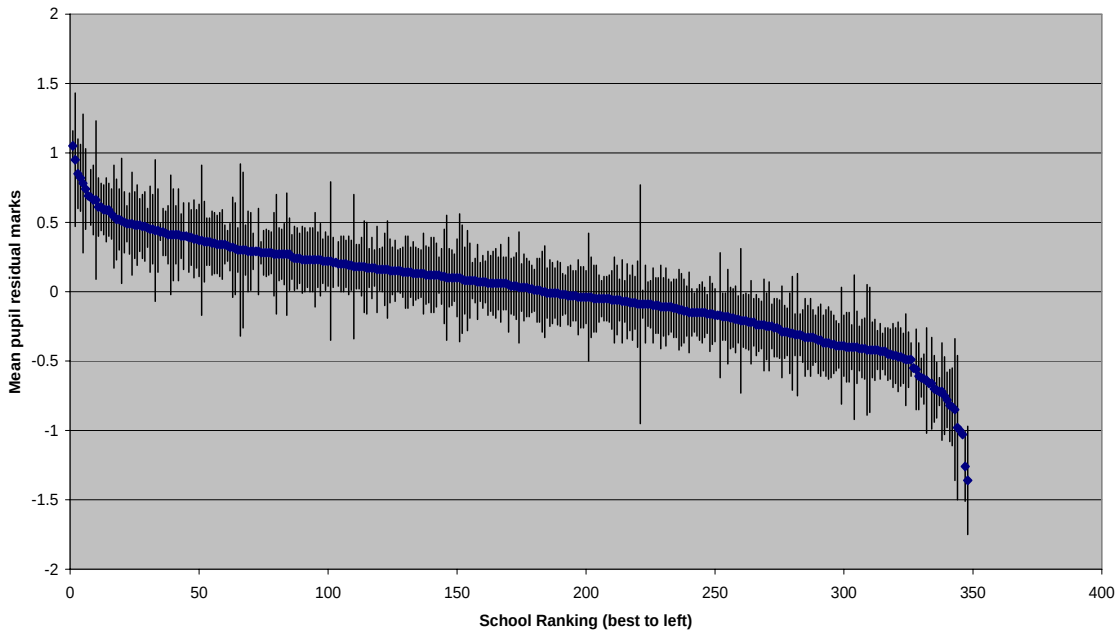


Figure A5.4 – School Ranking using Marks with 95% Confidence Intervals



Appendix 6 – Use of Multilevel Modelling

The same data used above were used in multilevel models. These models represent a sophisticated way of analysing school-based data and much research across the world has been directed at this approach over the last ten to twenty years. Unfortunately, the approach is specialist in its nature and this section is likely to be obscure to the uninitiated. Some readers may prefer to omit this section.

Multi-level models involve an extension of regression analysis in which the clustering of pupils within schools is explicitly recognised. They allow the analyst to partition the variance between pupils and schools and to run regression analyses at both of those levels explaining variance at the two levels. .

The initial analyses, shown below in the columns headed "Null" and "Model 2", indicate, as expected, that, as the key explanatory variable was added to the equation, about half (55%) of the variance at the pupil level was "explained". It was, however, surprising to see a reduction of only 14.5 per cent in the school level variance.

Table A6.1 - Multi-level models with KS2 APS as the outcome

	Null	Model 2	Model 3	Model 4
Cons	**28.53 (0.08)**	**12.39 (0.13)**	**19.56 (0.13)**	**23.15(0.61)**
KS1 APS		**0.570 (0.007)**	**0.570 (0..07**	**0.573 (0.007)**
Average deprivation			**-0.102 (0.019)**	**-0.153 (0.20)**
Average KS1 APS				**-0.227 0.038)**
Random				
School	**1.44 (0.16)**	**1.23 (0.12)**	**1.05 (0.11)**	**0.91 (0.094)**
Pupil	**9.70 (0.18)**	**4.38 (0.08)**	**4.39 (0.08)**	**4.39 (0.08)**
Explained				
School		14.50%	27.08%	36.81%
Pupil		54.80%	54.74%	54.74%
Proportion of variance associated with School	12.90%	21.90%	19.30%	17.17%

The outcome residuals of Model 2, the value-added scores, correlated 0.72 with those obtained from the null model, indicating an important difference in the rank ordering of schools and value-added scores as indicated by Model 2. However, when compared to the results calculated using a mean regression line, the correlation with the value-added scores was almost perfect at 0.994, indicating that the value-added scores derived from the multi-level model and from OLS regression were almost indistinguishable.

In Models 3 and 4 school level averages were added to the equations and this "explained" more of the school level variance. It also changed the rank orders of schools value-added scores as indicated by the correlations below.

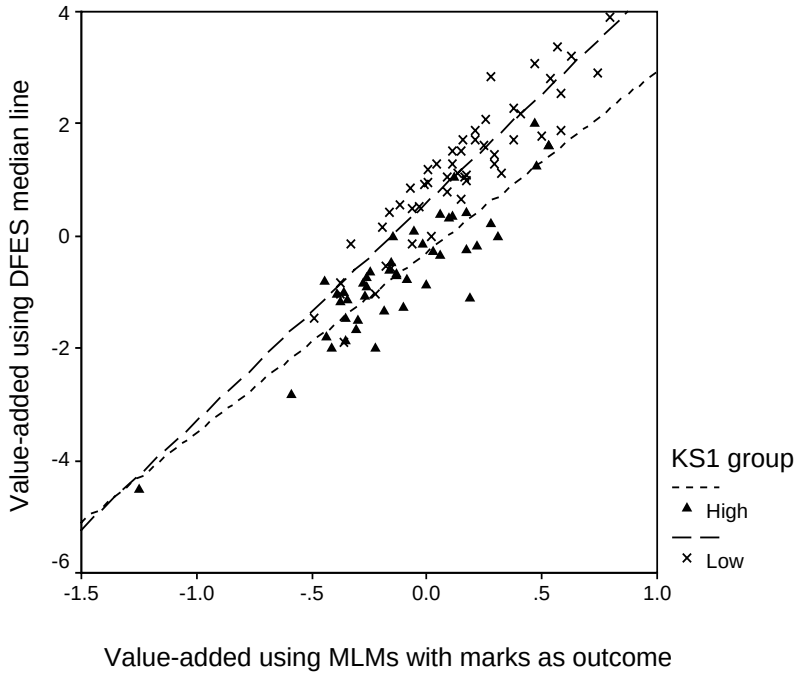
Table A6.2: Correlations between value-added scores for schools restricted to cohorts greater than 10 pupils

	DfES	Model 2	Model 3
Model 2	**0.96**		
Model 3	**0.93**	**0.96**	
Model 4	**0.85**	**0.91**	**0.95**

The correlation between the DfES value-added scores and Model 4, which takes into account the average deprivation and average KS1 scores of the schools, was 0.85. This suggests that some schools would appear in a very different light if the context of the school were taken into account.

It is possible to try alternative models using marks as the outcome, as before, and using the KS1 APS scores as the control. Now the value-added scores correlate with those produced by the median line at 0.905. The scattergram below shows the results from the multilevel analyses using marks plotted against the value-added using the median line. The results are plotted just for the highest and the lowest fifth of schools on the basis of their KS1 APS scores.

Figure A6.1 - Plot of DfES value-added scores against MLM value-added scores in which the outcome was marks



It is clear that the schools with the lower starting points, the lower Key Stage 1 results, were getting higher value-added scores using the DfES median line than those with the higher starting points when the same controls were employed. This underlines the points made earlier concerning the use of median scores.

To extend the analyses further, multi-variate multilevel models were constructed in which the three sets of outcomes at Key Stage 2, that is the mathematics, reading and science normalised marks, were put into a multilevel model as outcomes. This

appears in the table below and it shows that, in each of the three cases, between 13% to 18% of the variance was associated with the school, and that the pupil level scores correlated about 0.7 with one another for mathematics, reading and science, slightly less between mathematics and reading and at the school level that correlation was around 0.8 varying from 0.74 to 0.88. One would expect the school level correlations to be higher because they are aggregated results. Data were added in the form of marks from Key Stage 1, that is to say, the mathematics mark, the reading mark and the writing level converted into a normalised score.

Table A6.3 - Multivariate Null Model

	Maths	Reading	Science
Fixed			
Cons	0.01 (0.02)	0.02 (0.03)	0.00 (0.03)
Random			
School	**0.13 (0.01)**	**0.16 (0.02)**	**0.18 (0.02)**
Pupil	**0.86 (002)**	**0.86 (0.02)**	**0.83 (0.02)**
% Variance associated with school	13	16	18
Correlations			
Maths		0.77	0.88
Reading	0.62		0.74
Science	0.70	0.70	

(School above and pupils below the diagonal)

In the table below we can again see that a large proportion of the pupil level variance is “explained”. In the case of reading, 60%, science rather less at 44%, and mathematics at around 52%.

Table A6.4 - Multivariate Model 2

	Maths	Reading	Science
Fixed			
Cons	0.03 (0.02)	-0.10 (0.02)	-0.04 (0.03)
Math mark KS1	**0.58 (0.01)**	**0.16 (0.01)**	**0.36 (0.01)**
Reading mark KS1	**0.17 (0.01)**	**0.46 (0.01)**	**0.33 (0.01)**
Writing KS1	**0.03 (0.01)**	**0.19 (0.01)**	**0.05 (0.01)**
Random			
School	**0.16 (0.01)**	**0.16 (0.01)**	**0.17 (0.01)**
Pupil	**0.41 (0.01)**	**0.34 (0.01)**	**0.44 (0.01)**
% Variance associated with school	28	32	29
% Variance explained			
School	-23	0	5
Pupil	52	60	44
Correlations			
Maths		0.72	0.86
Reading	0.28		0.70
Science	0.43	0.41	

(School above and pupils below the diagonal)

The variance associated with the school was now quite high at around 30% in each of the cases. The correlations between the pupil level scores were now much less because the pupil level variance had been taken out of it. They were between 0.28 and 0.41, but the correlations between the schools were still fairly high, ranging from 0.7 to 0.86. However, the surprising thing in the models is the very low proportion of variance "explained" at the school level. This is surprising, because primary schools within England are largely segregated along the lines of housing estates. That is to say, in affluent areas the children tend to get higher levels and in the tougher areas children tend to get lower scores. When one takes into account the levels of pupils going to schools at an earlier stage, one expects the differences between the schools, which are so very apparent in the raw scores, to reduce. In the above model, in mathematics there was a bigger variance at the school level after controls had been put in than before, no change in the variance at all for reading, and a very slight reduction in science. Such results are unexpected and unheard of in school effectiveness studies, which are based on high quality objective data. The finding raises questions about the quality of the data at Key Stage 1. It is quite clear that the rank orders within schools are good because the pupil level scores are well explained, but it is not clear that they are so good when it comes to differences between the schools. If the marks systematically overrated pupils in tough schools whilst underrating in affluent schools, the kinds of results seen in the model might be obtained.

In the next model, school level averages were added, and the slopes associated with the KS1 maths in the prediction of maths at KS2 and the KS1 reading mark in the prediction of reading at KS2 were allowed to vary.

Table A6.5 - Multivariate Model 3

	Maths	Reading	Science
Fixed			
Cons	1.19 (0.22)	1.13 (0.22)	1.13 (0.22)
Math mark KS1	**0.59 (0.02)**	**0.15 (0.01)**	**0.36 (0.01)**
Reading mark KS1	**0.18 (0.01)**	**0.46 (0.01)**	**0.33 (0.01)**
Writing KS1	**0.03 (0.01)**	**0.20 (0.01)**	**0.05 (0.01)**
Average school KS1	**-0.07 (0.01)**	**-0.08 (0.01)**	**-0.06 (0.01)**
Average deprivation	**-0.033 (0.007)**	**-0.033 (0.007)**	**-0.046 (0.008)**
Random			
School			
Cons	**0.13 (0.01)**	**0.14 (0.01)**	**0.15 (0.01)**
Maths	**0.019 (0.003)**		
Reading		**0.013 (0.002)**	
Pupil	**0.39 (0.01)**	**0.33 (0.01)**	**0.44 (0.01)**
% Variance associated with school	25	30	23
% Variance explained			
School	0	13	17
Pupil	55	62	44
Correlations			
Intercepts			
Maths		0.68	0.84
Reading	0.29		0.66
Science	0.44	0.41	

(School above and pupils below the diagonal)
NB Correlations between slopes and between slopes and intercepts all <0.3

The first was the school average deprivation (derived from pupil level Jarman deprivation scores derived from Postcodes). This was a significant addition to the model for all three outcomes and it had a negative coefficient. This is because the greater the deprivation the less well the school was doing in its value-added score. Not a surprising result. When the average Key Stage 1 APS score was included it also proved to be a significant addition to the model, again with a negative coefficient. This suggests that the higher the school's average intake the lower the value-added score. That is a strange result because one expects and, indeed, commonly finds the opposite effect (see for example Harker and Tymms, in press). Again this suggests something odd with the data.

Now more of the school level variance is explained, but with figures from 0% to 17%, that is rather less than expected for this kind of study. One might well expect

something like 70% or even higher to be "explained" in a model of this sort and to find that none is explained in mathematics and only 17% in science is really odd.

The next point to note here is that the school level value-added scores still correlate strongly between mathematics and science (0.84) but the correlations between mathematics and reading is only 0.68, and between reading and science only 0.66. That suggests that in looking at the value-added scores of schools one should really be looking at the scores within curriculum areas and that to join them all together in an overall value-added score may be misleading.

Calculating an appropriate sample size

The Royal Statistical Society has called for the reporting of confidence intervals whenever performance indicators are published. Were this advice to be followed it might result in the more careful use of data. However, there is a danger that indicators may be misjudged even when accompanied by confidence intervals. With this in mind a question arises as to the appropriate sample size for which school based value-added measures should be published. In order to make a recommendation the reliability of value-added measures is considered.

Within a two level multi-level model (Goldstein 1997) of the type reported above the reliability is given by:

Formula I:
$$\text{Reliability} = \frac{n\sigma_u^2}{n\sigma_u^2 + \sigma_e^2}$$

Where
n is the number of pupils in the school sample
σ_u^2 is the school level variance
σ_e^2 is the pupil level variance

The proportion of variance associated with the school – the intra-class correlation (rho) is given by:

Formula II:
$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

A combination of the two formulae (I and II) allows for a calculation of the sample sizes associated with various values of rho. A series of graphs showing the connection were reported as part of the Value-Added National Project in Tymms and Henderson (1995) and again in Fitz-Gibbon (1997).

The value of rho for primary schools in England between the ends of KS1 and KS2 was estimated in Table A6.1 as 0.22. This piece of information can then be used with an appropriate reliability figure to estimate the minimum sample size for publication. In view of the accountability nature of published data it is appropriate to select a relatively high figure. When producing tests for the nation, QCA would surely be

aiming for reliabilities greater than 0.9, and it seems proper that schools based measures should be at least as reliable. A figure of 0.95 does not seem unreasonable.

For a reliability of 0.95 a sample size of 67 is needed.
For a reliability of 0.90 a sample size of 32 is needed.

We conclude that, assuming that appropriate procedures are used, a figure of at least 50 pupils on whom matched KS1 to KS2 data is available is needed before reasonably reliable value-added measures can be produced.

Appendix 7 – Reliability, Error and Bias

Confusion sometimes arises when a word like **reliability** is used in the context of value-added scores. If a measure is unreliable it does not mean that it has a mistake in it. Rather, it means that we must not put too much store by it as an indicator of what it purports to measure. And here we must be clear about what the published value-added measures imply. Implicitly they are taken as proxies for a primary school's effectiveness in teaching English, mathematics and science. This means something broader than success with the particular individuals who happen to have taken the KS2 assessments in that year. It suggests that if a different group of pupils had gone through Key Stage 2 at that school it would have got similar value-added results. Hypothetically we could imagine a school existing in a hundred parallel universes and having a hundred different value-added scores in each of those universes. The term reliability can be seen as the extent to which those hundred different results are consistent.

The other side of reliability is **error** and again the word can cause misunderstanding. In common day usage "error" means a mistake but in statistics "error" means the extent to which an indicator missed its mark – not through a mistake in the calculation or marking - but for some other reason, perhaps because the small number of pupils produced results which did not reflect the underlying performance of the school.

Finally, the word **bias** is used to indicate a systematic error – a tendency to miss the mark consistently in a particular direction.

Appendix 8 – Head Teachers' Conclusions about Value Added

Below are the conclusions some of the head teachers we interviewed have come to, having seen primary value-added league tables published nationally for the first time.

"I think parents have much more integrity than we give them credence for. My experience is that sometimes you'll get parents who are moving into the area or they have children who are starting in reception, who will have looked at the Ofsted, the League Tables, but they come in very pragmatically, saying we know SATs are not the be-all and end-all, we know you're a good school, we'd like to meet you, we'd like to look round to see the school in operation, and they tend to come with quite an open mind." (H1)

"If there is National Curriculum, if there is a Literacy and Numeracy Strategy, if there is Ofsted, if there is an LEA and consultants, and all the people who are monitoring - and governors – what's going on in schools, why do you need other measures that are not as effective?" (H3)

"A whole industry will now be generated on the back of this value-added process." (H4)

"I would ask them [the DfES] to look at what is happening in Wales and Scotland, and prove that standards are not rising in those two countries. And if they can't, I would ask them to do exactly what they've done there. So we would have a curriculum for 4 year olds to 7 year olds - consistent and continuous, topped by teacher assessment only. We would then go on to not publishing league tables at Key Stage 2 and having a mixture of a battery of standardised assessments available to teachers to support their teacher assessment judgements at the end of Key Stage 2." (H4)

"I really don't think people are that interested [in league tables] I think it was a one day wonder for a couple of years and I think parents now are much more astute and I think they collect a whole range of information when they're selecting schools." (H1)

"Our parents wouldn't send children here if they were too much influenced by those sort of things [league tables]." (H3)

"I think probably even the local press has become cynical now about it all … there's very little [local press coverage of league tables and value-added]." (H3)

"Parents who come here do tend to look at figures and things, but as yet they seem to be looking at the raw scores: they want to know that we're 11th, or that we've got 95% of children getting level 4's at the end of Key Stage 2, rather than value added, but whether that'll become more important - I think it's more staff that have been concerned, because when we've looked at figures we've said this looks as if we've done nothing with these children, and we've done an awful lot with them." (H7)

"I know from my previous school, we did well out of it, because we came in the top 5% nationally [value-added], and so it gives you something to look at, focus on and celebrate really apart from anything else, and I think it helps staff morale, because

you think gosh we are working hard and we have made a difference … yet equally the reverse of that is also true, where you do work your socks off and then you discover that you don't come out well in the value added." (H6)

"I think the whole thing's flawed … and I think they say how can we show improvement in schools that don't perform so well, and I've no problem against that, but if it's not equitable then it's not fair." (H8)

"Do away with league tables. I don't think they're helpful really … for any school. It doesn't make people work harder, I think. If the government think league tables are going to make people work harder … it just demoralises people because they think if they're going to rank us really low – we've worked so hard and it looks like we've done nothing - we wouldn't be much worse off if we didn't." (H7)

"The changes that I would want to see are really to do with the way in which the DfES and the support services, the local authority, actually support schools, rather than this whole climate of challenge. I'm not quite certain that the climate of challenge is what is needed …I worry about what the adversarial nature of it will do to schools and to the profession." (H5)

"The league tables have ceased to be a major talking point among heads and among parents really, unless there's a dramatic change in the results, and therefore if they're not noticing that they're hardly going to be looking at the extra data about value added." (H8)

"[Do parents and governors understand these tables?] No I don't think they do - they certainly don't understand value added." (H6)