Published article:
Homer, MS, Ryder, J and Donnelly, J (2011) *The use of national datasets to baseline science education reform: exploring value-added approaches.* International Journal of Research and Method in Education, 34 (3). 309 - 325 . ISSN 1743-727X

http://dx.doi.org/10.1080/1743727X.2011.609544

The use of national datasets to baseline science education reform: exploring value-added approaches

Matt Homer[1], Jim Ryder, and Jim Donnelly,

School of Education, University of Leeds

[1] Corresponding author. School of Education, University of Leeds, Leeds LS2 9JT

Summary

This paper uses data from the National Pupil Database to investigate the differences in 'performance' across the range of science courses available following the 2006 Key Stage 4 (KS4) science reforms in England. This is a value-added exploration (from KS3 to KS4) aimed not at the student or school level, but rather at that of the course. Different methodological approaches to carrying out such an analysis, ranging from simple non-contextualised techniques, to more complex fully contextualised multilevel models, are investigated and their limitations and benefits are evaluated. Important differences between courses are found in terms of the typical 'value' they add to the students studying them with particular Applied Science courses producing higher mean KS4 outcomes for the same KS3 level compared to other courses. The implications of the emergence of such differences, in a context where schools are judged to a great extent on their value-added performance, are discussed.

The relative importance of a variety of student characteristics in determining KS4 outcomes are also investigated. Substantive findings are that across all types of course science prior attainment at KS3, rather than that of mathematics or English, is the most important predictor of KS4 performance in science, and that students of lower socio-economic status consistently make less progress over KS4 than might be expected, despite prior attainment being accounted for in the modelling.

Introduction

Reforming the science curriculum

In 2006 a major reform of the science curriculum for 14-16 year olds (Key Stage 4) in England was enacted. The two main aspects of the reform were (i) the provision of a wider variety of science courses with more emphasis on applied science, and, (ii) more of a focus on teaching about the nature of science-, and socio-scientific issues. The reform was implemented partly as a result of a growing number of reports stating that the science curriculum should be made more relevant and useful to people in their day-to-day lives (Millar & Osborne 1998). An additional motivator for reform was the growing evidence that for the benefit of society as a whole, particularly economically, there needed to be an increase in the number of students taking up post-compulsory science (Department for Education and Skills 2005). An analysis of multiple aims associated with this reform is available elsewhere (Ryder and Banner 2010).

The work in this paper is part of a broader study examining the reaction of schools to recent changes in the school science curriculum in England. The Enactment and Impact of Science Education Reform (EISER) project [2] is a three year longitudinal study, jointly funded by The Economic and Social Research Council and the Gatsby Charitable Foundation, looking at the responses of schools, teachers and students to the 2006 reform. EISER has both quantitative and qualitative strands, and the work presented in this paper is focussed on aspects of the former of these. The qualitative research within EISER involves interviewing children and teachers in 19 broadly

---

[2] http://www.education.leeds.ac.uk/research/cssme/projects.php?project=99&page=1

representative schools over a three year period, with the aim of assessing the relationship between the intentions behind the reforms and how these played out in the classroom (Banner et al*.* 2009). [3]

A widening of provision and measures of success

One key area of interest in the impact of broadening of science options at age 14 is the degree to which there is evidence of differences between science courses in terms of improved student attainment over the course of Key Stage 4 (KS4). The curriculum change takes place in a context where, since 2007, a key government measure of school performance in England has been the percentage of KS4 students achieving two or more GCSEs at A*-C (or equivalent) in science. The extent to which school performance on this and other measures might be influenced by the specific suite of science course(s) that the school offers its students, and how such 'options' are taken up, is clearly of some interest, not least to students and schools, but more broadly to the scientific community, policy makers and society as a whole.

[3] This sample of schools is not claimed to be completely representative of all schools in England since this would be impossible in such a small sample, but rather aims to approximately reflect the profile of English schools in terms a variety of important educational characteristics. These include percentage of students who achieve 5 A*-C grades (or equivalent) at GCSE, religious affiliation, student ethnic mix, percentage of students eligible for free school meals, and awarding body in science at KS4.

The National Pupil Database and value-added studies

The National Pupil Database (NPD) is a longitudinal database containing student-level attainment data for all students in maintained schools in England. The assessment data is linked, via a unique student number, to a wide range of student characteristics collected in the Pupil Level Annual School Census (PLASC) including age, gender, ethnicity and measures of socio-economic status (SES). This combination of both attainment and social data allows not only the investigation into participation rates between different courses, but also an analysis of how much 'value' is added to students over the course of their studies, and, importantly, the extent to which these value-added measures might vary between different courses.

Previous value-added studies, often using multilevel modelling techniques, include the development of statistical models of varying complexity for estimating the longitudinal impact that teachers and schools have on their students. These studies generally use repeated measures on the students (often pre- and post-test scores) as the outcomes from which 'value-added' measures can be constructed (Thum 2003; McCaffrey et al. 2004). This research is often quite technical in nature, but more recently there has also been work with a more applied focus looking at, for example, measures of school effectiveness (Thomas et al. 2007). This latter study analysed trends in performance of a sample of English secondary schools over a ten year period, and investigated the extent to which there were discernable patterns in school improvement during that time. The 'value-added' by schools was measured by comparing a prior attainment measure, a test taken on entry to secondary school, with a measure based on final examination results students obtained at the end of secondary education.

There has been some criticism in the literature of particular value-added methodologies employed (particularly in the US) to evaluate the performance of school districts, schools and students (Amrein-Beardsley 2008). This paper also raises the important issue of missing data, something that shall also be discussed in relationship to the NPD in the current article.

A body of literature has also built up recently reporting specifically on the use of the NPD for investigating attainment-related aspects of performance in the maintained sector of English schools. These again include studies where prior attainment (hence 'value-added') is accounted for in the modelling. Examples of such work include Schagen and Schagen's study (2005) looking at the effects of different types of school on students' progress. A recent paper (Noyes 2009) used the NPD to explore how the completion of A-level mathematics varied by student characteristics and GCSE mathematics grades. However, there is little evidence in the literature of the use of the NPD to carry out a direct and systematic comparison of a suite of related courses on a value-added basis. For school science education in England, where the 2006 reforms have increased the number of different courses (and types of course), there is a particular richness to the ecology of KS4 qualifications that makes such an investigation compelling. In addition, the importance of science, and of a scientifically literate population to the future progress of society, is widely acknowledged (Roberts 2007)

This paper

Using the data from the NPD, this paper investigates the variation in the patterns of attainment across a set of distinct science courses in England. The analysis is carried out on a value-added basis, taking into account prior attainment at the end of Key Stage 3 (KS3, age 11-14). Three distinct statistical methodologies, with increasing levels of complexity, are employed. The overall aim of the paper is to compare and contrast the three statistical methodologies in terms of the quality of the substantive findings that they provide, and of the limitations and benefits that they each bring. Their usefulness in terms of the interpretability of the results to, for example, policy makers and other stakeholders will be considered. The nature of the inferences that can justifiably be drawn from such research will also be discussed.

Methods

The first set of students involved in taking the post-reform courses began Year 10 in 2006, and took their KS4 examinations and assessments two years later in the summer of 2008. The findings in this paper are for this first post-reform cohort of students. In order to simplify the analysis, students were grouped into one of six categories of science ‘course’ based on the qualifications they achieved in the 2008 examinations. This breakdown is shown in Table 1, where participation by gender for each course is also shown. There is an additional, seventh, category made up of those students in the cohort who, whilst present in the Pupil Annual Level Census (PLASC) data, were not present in the NPD under any of the six science courses detailed above. Since the studying of science is a statutory requirement at KS4, it is likely that the majority of students in this last group did follow a science course throughout KS4 but were not entered for any examination/assessment in 2008.

The courses in Table 1 might be interpreted as arranged broadly in descending order of the emphasis on traditional scientific knowledge. For example, Triple Award Science or Dual Award Science provide the standard routes onto separate science courses at post-16 level (e.g. Biology, Chemistry and Physics A-levels). Applied sciences courses have a stronger focus on the use of science knowledge within vocational contexts such as healthcare and forensic science. Entry Level Qualifications (ELQs) are targeted at students working at below grade G at GCSE level, whereas courses in the other five groups cover the full range of GCSE (or equivalent) grades. For this reason ELQs will not be studied further in this paper.

The gender breakdown gives an indication of the variation in types of students who achieve qualifications in particular courses. For example, for Triple Award Science females are underrepresented compared to males, whereas for Other Applied Science they are overrepresented. The article by Banner and colleagues (Banner et al*.* forthcoming) provides more details on how participation in these courses varies in a number of other important ways, including by socio-economic status.

Table 1 shows that the modal science course is Dual Award Science, taken by over 50% of the cohort. However, the proportion taking this course is in decline over time, and that of some of the others is increasing, particularly Triple Award, and the Applied courses (Banner et al*.* forthcoming). However, the purpose of this paper is not to investigate longitudinal change, but rather to provide an early comparison of these courses, in value-added terms, for the first cohort following them after the enactment of the reform.

TABLE 1 HERE

Three variants of value-added comparisons between courses will be presented as follows:

1. Simple value- added – the mean KS4 attainment by KS3 science level.

2. Regression value-added – ordinary least squares regression of mean KS4 attainment on KS3 science fine level. Fine level scores are based on the interpolation of discrete test scores between key stage levels, and provide a more fine grained measure of KS3 performance.

3. Contextualised value-added – multilevel modelling analysis to take account of students clustering in schools, with students as level 1, schools as level 2 and including a wide range of predictors in the model such as KS3 fine levels in science, mathematics and English, as well as a variety of student background characteristics.

In the NPD, the outcome at KS4 is recorded for GCSEs as a points score on a scale from 16 points for grade G, to 58 points for A*, with the difference between successive grades worth six points. For non-GCSE courses, the level of the award is translated to the GCSE equivalent number of points so that all courses are scored on the same scale.

In order to make direct comparisons between courses, the mean attainment across the separate GCSEs making up the course is used as the measure of KS4 student outcome in science within each course. So for example, for Triple Award this is the mean of the separate Biology, Chemistry and Physics GCSEs, and for Dual Award Science this is the mean of Core Science and Additional Science.[4] Hence, in the results that follow, the comparison between courses is one based on the quality of outcomes, rather than the quantity. This implies that any differences in value-added attainment that might be found automatically scale up, in terms of total GCSE points, when comparing courses that are worth a different number of GCSEs.In terms of ‘size’, all of the first four courses listed in Table 1 are worth at least two GCSEs, with Triple Award worth three, and Other Applied courses worth two or four.

Results

Method 1: Simple value-added

For each of the five science courses identified, Figure 1 shows a graph of mean KS4 points against KS3 science level.[5] It should be pointed out that in many ways this provides for a crude comparison of courses in terms of value-added for a number of reasons including:

(i) It takes no account of the variation in types of typical student profiles following each of these courses. It is well known (Gorard & See 2009), that

[4] For readability and to avoid confusion when group means are compared, these mean KS4 outcomes will be referred to throughout simply as ‘KS4 points’.

[5] The full figures are given in Table A1 in the appendix. As a result of the large sample sizes, all 95% confidence intervals for the true mean KS4 performance in Figure 1 would have a half-width less than 1 KS4 point.

participation (and attainment) across science courses is highly stratified by, for example, socio-economic status.

(ii) The value-added comparison is based only on one prior-attainment measure, that of KS3 science. This KS3 measure is itself limited in terms of the discrimination it affords between students, with the majority of the cohort (87.4%) being awarded between levels 3 and 6.[6]

(iii) The assessment regimes across each of these five courses are diverse, with some being assessed largely through formal examination (e.g. Triple Award), whereas others (e.g. Other Applied) involve a large proportion of coursework in their assessment.

(iv) The allocation of specialist teachers, the timetabled time allocated to each course, and the type(s) of pedagogy employed might vary considerably between the courses.

FIGURE 1 HERE

Despite the limitations of this approach, some obvious differences in patterns of performance emerge. The most striking finding is that at the lower end of KS3 range (levels 3 and 4), the Other Applied Science courses provide much greater value-added in terms of KS4 outcomes for the same KS3 science level. The difference is

[6] In later, more complex, methods, KS3 fine levels are used as prior attainment measures. However, Method 1 is intended as a simple and intuitive, yet useful, approach with parsimony at a premium. Hence, KS3 main levels are employed here.

7.4 points, at level 4 compared to the Dual Award Science (i.e. of the order of a whole GCSE grade). However, at the other end of the KS3 range the difference between these two courses is in the other direction, with Other Applied Science 1.4 KS4 points below Dual Award Science at KS3 level 6. There is similar, although smaller in magnitude, cross-over between Dual Award Applied and GCSE Science Only. Such cross-over patterns in performance when comparing between ‘applied’- and ‘non-applied’-type science courses has been evidenced before (Bell et al. 2009) where (pre-reform) GCSE courses in Applied Science and Double Award Science were compared on a value-added basis (KS2 to KS4).

Other patterns that can be observed are that Triple Award Science and Dual Award Science tend to converge as KS3 levels increase, whilst Dual Award Science and Dual Award Applied Science diverge. However, Other Applied Science excepted, all four course follow a generally linear trajectory, with similar slopes. Triple Award shows the most value-added (for a fixed KS3 level), usually followed by Dual Award Science, with Dual Award Applied Science and GCSE Science Only generally showing the least value-added.

Method 2: Regression-based value-added

The degree of linearity of the graphs shown in Figure 1 implies that an ordinary least square (OLS) regression-based approach, using KS4 science outcomes predicted by KS3 (fine) science levels, is likely to result in broadly similar patterns of performance, since under regression linearity is assumed. However, KS3 fine levels are, in theory at least, a more accurate and discriminating measure of KS3 performance compared to (whole) levels as used in Method 1 so there is scope for differences to emerge. The model is given formally, for the *i*th student, by

$$(KS4\_Points)_i = a + b(KS3\_ScienceFineLevel)_i + e_i$$

where e_i is the student residual (these assumed to be normally distributed and independent of each other). The regression parameters, *a* and *b*, are estimated using traditional OLS methods.

Figure 2, a plot of the predicted regression lines for each course, shows that under such a model, Other Applied Science again provides much greater value-added at the lower end of the KS3 scale, but the opposite at the upper end. The full figures are given in Table A2 in the appendix. The other courses show some differences between each other depending upon exactly where on the horizontal scale one makes the comparison. All four courses are fairly close together at the lower end of the KS3 scale but separate out into two distinct pairs (Triple Award Science and Dual Award Science, and Dual Award Science Applied and GCSE Science Only) further up the scale.

Considering these pairs of courses in turn, there is a small degree of crossover between Triple Award Science and Dual Award Science, with the latter adding more value at the lower end of the KS3 scale, but generally these two courses are closely aligned. Similarly, Dual Award Applied Science, whilst having a little more value-added at the lower end of the KS3 scale, is broadly similar to GCSE Science Only across the range.[7]

FIGURE 2 HERE

It should be remembered that the limitations of Method 1 in terms of providing a robust and meaningful comparison between science courses also apply here. More will be said on this issue in the discussion.

Method 3: Contextualised value-added with multilevel modelling

The OLS regression modelling employed in Method 2 is flawed in the sense that it assumes that students clustered in schools perform independently of each other, a basic assumption of OLS methods, whereas this is known not to be the case; students' performance tends to be correlated with that of fellow students in their school. This is the key reason why it was decided to not make Method 2 more complex through the addition of additional predictors. Ignoring the dependency in the data can give rise to misleading results, and more sophisticated techniques that take account of the hierarchical structure of the data need to be employed (Dorman 2008), although this had been disputed by some researchers (Gorard 2007; Gorard 2009). Multilevel methods can, however, provide better insights into the data than

[7] It should be noted that a direct comparison between Figures 1 and 2 is not straightforward since in the former, a student awarded, say, level 4 might achieve a fine level anywhere between 4 and 5. Hence, in order to carry out such a comparison, the horizontal scale in Figure 1 should be relabelled by adding 0.5 to each of the existing number labels.

are possible with traditional regression techniques. For example, the variation by school (i.e. the 'school effect') can be measured across each course as a whole, and at the individual school level (through school residuals), thereby giving indications of the differences between schools in terms of their value-added 'performance'.

The simplest multilevel model (assuming a variance components structure) is given, for each course, in the usual way by:

$$(KS4\_Points)_{ij} = a + b(KS3\_ScienceFineLevel)_{ij} + u_j + e_{ij}$$

In this equation, subscripts i and j refer to student and school respectively, a and b are constants, u_j is the school (level 2) residual, and e_{ij} is the student (level 1) residual. Both residuals are assumed to be normally distributed and independent of each other.

Since part of the aim of this research is to provide a telling description of what is influencing KS4 performance differentially across the five science courses, in addition to employing a more appropriate statistical methodology, there is also the need to contextualise the model(s) to take account of the variation in the types of students doing each of the courses. The PLASC data within the NPD provides a wealth of data at the pupil level, including the familiar variables of age, gender, and ethnicity, as well as many other measures of student background including socio-economic status.[8] Other predictors of KS4 performance that are likely to be of import include KS3 mathematics and English attainment (fine levels). Decisions as to which variables to include amongst the many available are always contestable (again, see the discussion for more on this).

The equation above is therefore extended in the natural way to add additional predictors. Table 2 shows the results of five separate two-level (students nested in schools) main effects models of KS4 outcomes for each of our five science courses using the same set of predictors in each (KS3 attainment, SES, gender, age and ethnicity).[9] MLwiN software was used for all these analyses (Rasbash et al. 2009).

Unfortunately, it is not possible to produce a meaningful graph for these models corresponding to those given in Figures 1 and 2 for the earlier methods, and this could be seen as a negative feature of this method. Other than those starred, the coefficients shown in Table 2 are all significantly different from zero (at the 5% level) but for reasons of space standard errors and/or p-values have been omitted. However, as in all educational research, a distinction should be made between statistical and educational significance. With large datasets, even small and practically unimportant differences can be statistically significant. Care also needs to

[8] The Index of Deprivation Affecting Children (IDACI) measures the proportion of children under 16 in the area where the student lives who are living in an income-deprived family. The free school meals (FSM) variable is binary, indicating whether or not a student is eligible to receive free school meals.

[9] We have basically assumed that the regression 'lines' for each school can vary, but only in the sense that their intercepts on the vertical (KS4) axis can be different (whilst collectively following a normal distribution). In every other sense the 'lines' for each school are the same (i.e. are parallel).

be taken when looking at the absolute size of the predictors since they are not all on the same scale. For example, IDACI takes values in the range 0 to 1, and hence a unit change in this measure is equal to the length of the whole scale. Variables could have been standardised but to aid interpretation have been left unstandardised.

The first thing of note in Table 2 is the wide disparity between the constant terms across the courses. This is the value of the outcome if all the predictors are zero, and, in this context, needs to be interpreted carefully. However, the relatively large value of this for the Other Applied Science course (19.49) does indicate that KS4 outcomes start from a higher base compared to the other four courses. This corroborates the earlier findings (Figures 1 and 2) where students from this course had greater KS4 points for the same KS3 level, especially at the lower end of the scale, compared to the students doing any of the other four courses.

It is clear that KS3 attainment is an important predictor of KS4 attainment, and that (within course) KS3 science is always the most important predictor. So, for example, for Triple Award Science, a unit change in KS3 fine level implies an increase in KS4 points of 4.91 (*assuming all other predictors stay the same*). However, the relative positions of English and mathematics within different courses show some important differences. In Triple Award Science, the mathematics coefficient (2.48) is almost double that of the English (1.37), presumably reflecting the relative importance of these skills in influencing KS4 performance. By contrast, for Dual Award Applied Science, the English coefficient (1.92) is larger than is that for mathematics (1.50), suggesting that English skills are more important in influencing outcomes on this course, compared to mathematics.

Moving on from the prior attainment variables, it is clear that the coefficients for gender are quite small, with some positive and some negative. The socio-economic status coefficients for FSM and IDACI are negative in all the courses, indicating that students with lower SES (but in other ways exactly the same) tend to perform worse than might otherwise be expected. Again the effects are not particularly large. Increased age has, perhaps slightly surprisingly, a small but consistent negative effect on performance.

The results for ethnicity are revealing, in that they are broadly positive indicating that in most cases, having controlled for all other predictors, ethnic minority students perform better than their white British counterparts (the reference group in each model). Whilst on the face of it these findings indicate the positive effect of being a member of an ethnic minority group on (value-added) performance this is, to some extent, an artefact of the inclusion in the model of prior attainment, which is itself highly stratified by ethnicity (Department for Children, School and Families 2006). The same research has also shown that socio-economic status is stratified by ethnicity. The complex interplay of different student characteristics underlines the potential danger in analysing educational outcomes using models that are too oversimplified. If the interactions or confounding amongst particular variables is not sufficiently accounted for then incorrect inferences can easily be drawn.

The final section of the table gives model summary data, including the final sample sizes at the student and school levels. The residual school effect gives an indication of how much schools vary in their ‘performance’ for each course. It estimates the percentage of the variation in KS4 attainment that can be attributed to the schools rather than the students, once variation in other factors (i.e. the predictors in the model) has been allowed for. Hence it is clear that, for example, schools vary much more in their ‘performance’ in Other Applied Science (46.9%) than they do in Dual Award Science (17.1%). The general pattern is that the more established courses appear to show a smaller school effect, and that Other Applied Science courses form an outlier group in this regard. It will be important to see in the future whether or not there is a degree of levelling out in these apparent differences across courses as the reform takes hold in schools. Note that within a school, a particular science course usually corresponds to a single awarding body specification for the course, meaning that the ‘school’ effect might be more accurately thought of as a ‘specification within school’ effect.[10] Hence, apparently large ‘school’ effects might be due, in part, to variation at the specification level, rather than that of the school. Probing of the role specifications in influencing course ‘performance’ might form part of future research.

TABLE 2 HERE

Listed earlier there were four limitations of Methods 1 (and 2) when it comes to making valid course comparisons. Of these, the first two, concerning the simplicity of the model in terms of predictors, do not apply to Method 3. However, the last two, related to the variation in assessment regimes and other key differences between the courses that are not accounted for in the modelling, certainly do.

---

[10] Courses can have multiple specifications. For example, for Triple Award courses each of the three main awarding bodies in England provides at least one specification.

Students changing schools

A comparison within the NPD of school codes at KS3 and KS4 indicated that approximately 7.7% of students within the dataset had changed schools over this key stage. Further, an increasing proportion of students moving school was observed as one moves down the courses in Table 1, from Triple Award (4.1%) through to ELQ courses (13.9%). In light of recent research (Leckie 2009) which showed a negative relationship between student mobility and achievement, this might be expected to influence the findings of this section, particularly concerning the 'school' effect. For this reason, the results of the analysis have been compared with a parallel analysis carried out on the sub-sample of students whose schools were unchanged from KS3 to KS4. The substantive findings shown in Table 2, including the size of the 'school' effect, match those of this sub-analysis. It was therefore decided to present the results for all students, including those who had moved schools, in order to ensure that for each course the same maximal sample could be used across the three separate value-added methods under consideration.

Missing data

Approximately 6.6% of cases are missing from this multilevel analysis compared to that of the full cohort shown in Table 1. This is mainly due to missing KS3 fine levels in the NPD, with the majority of students missing KS3 results through either absence or not operating at the level of the test. The extent of this missing data varies across courses (from 3.4% for Triple Award, to 10.9% for the Single GCSE in science) and is therefore unlikely to be missing at random. However, for the other predictors employed in Method 3, the extent of missing data is small with, for example, the IDACI measure present in 99.6% of cases. None the less, an element of caution should be exercised in interpreting the results for this, and to a (slightly) lesser extent, the earlier, less complex models.

Discussion

There are two overlapping foci in this paper, a study of the differing patterns of value-added attainment of KS4 science courses following the 2006 reforms, and, within this context, a critique of possible methodologies that might be employed in order to make such comparisons. The discussion begins with a brief comparison of the *substantive* findings between the methods. It then goes on to consider the *methodological* differences between them, and to discuss the limitations and benefits of these types of analysis in terms of what conclusions can and cannot be justifiably drawn from them.

The ‘performance’ of courses across the different methods

The patterns of attainment resulting from the three methods employed in this analysis have some common features across the five KS4 science courses investigated. Across all three methods, students at the lower end of the KS3 range studying Other Applied Science perform considerably better at KS4 than might be expected on the basis of their KS3 grades alone. One interpretation might be that these Other Applied Science courses are not as closely related to KS3 science as the others are. It should be remembered that these courses are equivalent, in points terms, to between two and four single GCSEs, depending on the exact details of the specification, and so the ‘additional’ benefit that some students receive is actually magnified compared to the majority of the other courses being compared. These four remaining courses (Triple Award, Dual Award, Dual Award Applied, and GCSE

Science only) are broadly similar (within method) in terms of the importance of KS3 science levels in influencing their outcomes.

Across methods, it is clear that findings for Methods 1 and 2 are broadly similar, and this is largely due to the approximately linear graphs observed in Method 1. Whilst Methods 1 and 2 can be used to answer exactly the same research questions, direct comparisons between the results of these two methods and those of Method 3 are not possible in any straightforward sense. It is clear that the sophistication of Method 3 means that a broader set of issues can be investigated.[11] Some of the key substantive findings for policy under this latter method include the following:

- Across all types of course, KS3 science is the most important predictor of KS4 outcomes. The relative importance of KS3 mathematics and English in influencing KS4 outcomes varies by course.
- Students with lower SES tend to have lower KS4 outcomes, despite controlling for prior attainment. In other words, such students generally make less progress over KS4 than might otherwise be expected.
- Ethnic minority students tend to make greater progress over KS4, in value-added terms, than do their white British counterparts.

---

[11] It is possible to add additional predictors to Method 2, and that then some of the substantive findings might be similar. However, without accounting for the hierarchical structure inferential errors are possible.

- The role of schools in influencing KS4 outcomes varies greatly across the five courses.

A caveat on over-interpretation of findings

Findings on course differences in 'performance' must be treated cautiously and not be seen as promoting one type of course over another. There is no evidence in this study that changing an individual's course based on the analysis presented would automatically lead to a change in this student's KS4 performance. Evidence is presented elsewhere (Banner et al*.* forthcoming) of the different stratifications that occur in terms of participation in these courses by, for example, KS3 attainment, ethnicity and SES. These differences imply that the outcome of changing from one course to another cannot be predicted on the basis of the results presented in this paper. However, there are indications in the findings that two hypothetically identical students (in the sense of the predictors used here) might have different expected (i.e. average) KS4 science outcomes based on the course they chose.

Methodological issues

There have been criticisms of some applications of value-added approaches, particularly as employed by the UK government to produce school league tables (Gorard 2006; Leckie & Goldstein 2009). These criticisms are not, *per se*, about the methodology employed, but rather the incorrect or misleading uses that the findings of the analysis can be put in ranking schools by performance, or to inform school choice. Whilst any limitations of the applications of the findings have been plainly stated, it is clear that value-added approaches are indispensible tools for this type of data. However, when carrying out value-added analyses, there can be a tension between using simple, but more intuitive approaches, compared with more complex techniques that require careful mediation by experts, as well as serious investments of time and effort on the part of non-specialists in order to fully understand the meaning of the results. Comparing the three value-added methods with this tension in mind, it is clear that Method 1 is the most intuitive, and provides simple, though arguably crude, insights into the differences between courses in terms of the relationship between KS3 and KS4 performance in science.

Through the use of KS3 fine levels as opposed to main levels, Method 2 has the apparent advantage over Method 1 of using a more discriminating measure of KS3 performance. However, the difference in quality of these two KS3 measures should not be overstated. Beyond this, the second method also provides a measure of the degree of linearity in the relationship between science attainment at the two key stages through the r-squared measure (Table A2), although the downside of this is the necessary assumption in the regression that the underlying relationship is linear, a condition not required under Method 1.

Overall then, parsimony would suggest that Method 1 holds the edge over Method 2 in terms of the simplicity of the method and the ease of understanding and interpretability of the findings. The limited additional insights afforded by the more complex method (2) do not seem ‘worth’ the extra degree of sophistication of the method, certainly in this context.

Quantitative researchers would generally agree that Method 3 provides the most ‘correct’ approach to such an analysis, although there has been some debate about the extent to which multilevel modelling is ‘worth it ’ compared to OLS-based techniques (Gorard 2007; Hutchison & Schagen 2008; Gorard 2009). Few would dispute that this methodology does offer clear benefits over the other two, in the sense that it allows for the relative importance in the modelling of a wide range of predictors to be measured, whilst also taking account of students being nested in schools. However, there are also negative issues more complex methods bring that are not easily resolved. Most obviously, the substantive findings are harder to interpret and understand, particularly as simple visual interpretations are not possible.

However, there is a further problem, one that is especially acute when employing data from the NPD. The final set of variables chosen to include under Method 3 as predictors is always arbitrary since with the NPD there is an almost endless supply of potential variables that could be included (Noyes 2009). As the main interest is in comparing across courses, it was decided in advance to use the same set of predictors within each course to make the comparisons. Previous research (for example Noyes 2009; Leckie 2009), and the qualitative strand of the EISER project provides a useful guide in this selection, but there remains no objective way of privileging this particular set of predictors from amongst all those available. What of, for example, a student’s special education needs status, what of school level predictors such as percentage of students eligible for free school meals, and what about possible interactions between any of the predictors? What of non-linear effects (modelled by included squared, and possibly, other higher order predictor terms)? None of these additional factors have been accounted for in the modelling, and it would always be possible to choose a different, arguably ‘better’ model.

Beyond the choice of what variables to include in the modelling, there are additional methodological issues that arise including the treatment of particular predictors as interval (e.g. IDACI). Even the outcome variable (KS4 attainment) is not properly interval but has been assumed to be so for the purposes of the analysis (this particular problem applies to Methods 2 and 3). Furthermore, the (random intercepts) multilevel model employed here is the simplest of those two-level multilevel models that could have been chosen. For example, the complexity of the modelling could have been increased by allowing the coefficients, in particular KS3 science attainment, to vary by school in a random slopes model. In fact, a limited additional analysis of this kind was carried out (not presented here), and the models did improve, implying that there is statistically significant variation in the coefficient for KS3 fine level science attainment across schools. Essentially, within a course, the rate of value-added based on KS3 science fine level is not the same across all schools.[12] Again, this is likely to prove difficult to report and explain in practical terms to non-specialists.

Finally, it is worth pointing out that statistical models employing greater numbers of predictors are more likely to suffer from other problems, such as non-random missing data, although this does not appear to be a major problem in this study.

The point of this part of the discussion is not to undermine the multilevel methodology employed, or its application in this particular science education context.

---

[12] However, the other fixed effect coefficients were not substantially different from those presented.

Indeed, the results obtained are meaningful, robust and give useful and important insights into the relative ‘performance’ of each of the five science courses being investigated. The critical point to emphasize is that there is no single model/methodology that is the best for all occasions or audiences, and that it is important that researchers are clear about this. More complex methods generally bring benefits, when employed appropriately, but usually they also raise problems with regard to communicating findings meaningfully to stakeholders who do not have specialist knowledge.

Conclusion

Using a variety of methodological approaches, this work has given descriptions at the aggregate level of patterns of KS4 science attainment following important curricula reform in England. Any influence on policy that these findings might have should, however, only occur at the national level, and not at that of the school or student levels. For example, there are dangers in a crude reductionist approach that states that this course is ‘easier’ than the other, and that therefore students should be encouraged to take the ‘easier’ option. The type of modelling carried out here, whilst insightful in many regards, remains of limited use in terms of providing a complete understanding of how the system is really working. It should always be remembered, notwithstanding the hundreds of variables available in national datasets, that many of the most important aspects of what actually goes on in schools are either difficult or impossible to measure, or are not included in the NPD. Further targeted work, including qualitative study, is needed to complement the findings in this paper in order to provide a more comprehensive view of how the reforms are working through in schools.

References

Amrein-Beardsley, A., 2008. Methodological Concerns About the Education Value-Added Assessment System. *Educational Researcher*, 37(2), 65-75.

Banner, I., Ryder, J. and Donnelly, J., 2009. The role of teachers' priorities for science education in the enactment of science curriculum reform. Paper presented at the European Science Education Research Association conference, 31st August – September 4th, Istanbul, Turkey.

Banner, I., Donnelly, J., Homer, M. and Ryder, J., Forthcoming. The impact of recent reforms in the KS4 science curriculum. *School Science Review*.

Bell, J., Donnelly, J., Homer, M., and Pell, G., 2009. A value-added study of the impact of science curriculum reform using the National Pupil Database. *British Educational Research Journal*, 35, no. 1, 119-35.

Department for Education and Skills, 2005. 14 - 19 Education and Skills. London: HMSO.

Department for Children, School and Families, 2006. Ethnicity and Education: The Evidence on Minority Ethnic Pupils aged 5-16. DCSF Topic Paper 0208-2006 DOM-EN.

Dorman, J.P., 2008. Conducting statistical tests with data from clustered school samples. *International Journal of Research & Method in Education*, 31 no. 2, 113-24.

Gorard, S., 2006. Value-added is of little value. *Journal of Education Policy*, 21 no. 2, 235-243.

Gorard, S., 2007. The dubious benefits of multi-level modeling. *International Journal of Research & Method in Education*, 30 no. 2, 221-36.

Gorard, S., 2009. Misunderstanding and misrepresentation: a reply to Hutchison and Schagen. *International Journal of Research & Method in Education*, 32 no. 1, 3-12.

Gorard, S. and See, B.H., 2009. The impact of socio-economic status on participation and attainment in science. *Studies in Science Education*, 45 no. 1, 93.

Hutchison, D. and Schagen, I., 2008. Concorde and discord: the art of multilevel modelling. *International Journal of Research & Method in Education*, 31 no. 1, 11-8.

Leckie, G., 2009. The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172 no. 3, 537-554.

Leckie, G. and Goldstein, H., 2009. The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172 no. 4, 835-851.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A. and Hamilton L. 2004. Models for Value-Added Modeling of Teacher Effects. *Journal of educational and behavioral statistics : a quarterly publication sponsored by the American Educational Research Association and the American Statistical Association*, 29(1), 67-101.

Millar, R., and Osborne, J. (Eds.), (1998). *Beyond 2000: Science education for the future*. King's College London, School of Education.

Noyes, A., 2009. Exploring social patterns of participation in university-entrance level mathematics in England. *Research in Mathematics Education*, 11 no. 2, 167-183.

Rasbash, J., Steele, F., Browne, W.J. and Goldstein, H. 2009. *A user's guide to MLwiN Version 2.10*. Centre for Multilevel Modelling, University of Bristol

Roberts, D. A. (2007). *Scientific literacy/science literacy*. In S. K. Abell & N. G. Lederman (Eds.), Handbook of Research on Science Education. Mahwah, NJ: Lawrence Erlbaum Associates.

Ryder, J and Banner, I. 2010. *Multiple Aims in the Development of a Major Reform of the National Curriculum for Science in England*, International Journal of Science Education,, First published on: 16 July 2010 (iFirst).

Schagen, I. and Schagen, S. (2005). Combining Multilevel Analysis with National Value-Added Data Sets: A Case Study to Explore the Effects of School Diversity, *British Educational Research Journal*, Vol. 31 no. 3, 309-28

Thomas, S., Peng, W.J. and Gray, J., 2007. Modelling patterns of improvement over time: value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education*, 33(3), 261.

Thum, Y.M., 2003. Measuring Progress Toward a Goal: Estimating Teacher Productivity Using a Multivariate Multilevel Model for Value-Added Analysis. *Sociological Methods Research*, 32 no. 2, 153-207.

Appendix

TABLE A1 HERE

TABLE A2 HERE

Captions

Table 1: 2008 KS4 participation in science

Figure 1: 2008 Mean KS4 examination points by KS3 science level

Figure 2: Regression of 2008 KS4 examination points by KS3 science fine level

Table 2: Multilevel modelling regression coefficients for KS4 points

Table A1: 2008 KS4 examination points by KS3 science level

Table A2: Regression of 2008 KS4 examination points by KS3 science fine level

Table 1

Course[13]	Overall student numbers	% of full cohort	Male		Female	
			N	% within male	N	% within female
1. Triple Award Science[14]	51,079	8.4	29,125	9.3	21,954	7.4
2. Dual Award Science[15]	333,080	54.7	167,002	53.6	166,078	55.8
3. Dual Award Applied Science[16]	50,539	8.3	23,802	7.6	26,737	9.0
4. Other Applied Science[17]	40,090	6.6	18,347	5.9	21,743	7.3
5. GCSE Science Only[18]	85,826	14.1	44,183	14.2	41,643	14.0
6. Entry Level Science Qualification	10,664	1.8	6,504	2.1	4,160	1.4
7. None of the above science courses	41,001	6.7	24,144	7.7	16,857	5.7
TOTAL	**612,279**	**100.5**	**313,107**	**100.5**	**299,172**	**100.5**

[13] Students can appear more than once across the courses, although the numbers doing so are very small (much less than 1%). However, within each course each student is unique.

[14] These are the students entered for all three of the separate GCSE science courses in Biology, Chemistry and Physics.

[15] Students entered for GCSE Additional Science (these students will also have been entered for GCSE Science)

[16] Students entered for GCSE Additional Applied Science (these students will also have been entered for GCSE Science)

[17] Students entered for Dual Award Applied Science, BTEC First Diplomas, OCR Nationals.

[18] Students entered for GCSE Science and no other course (other than approximately 4,000 doing this and Entry Level Science)

Table 2

Predictors (fixed effects)		KS4 science course				
		Triple award	**Dual award**	**Dual award applied**	**Other applied**	**GCSE Science only**
Constant		-3.24	0.07*	5.43	19.49	2.48
KS3 fine level	Science	4.91	4.83	3.72	1.76	4.26
	Mathematics	2.48	2.06	1.50	0.95	1.58
	English	1.37	1.41	1.92	1.24	1.67
Gender (female)		0.22	-0.06	0.69	0.73	-0.19
Socio-economic status	FSM	-0.31	-0.66	-0.92	-0.70	-1.08
	IDACI	-1.91	-2.07	-2.16	-1.37	-2.72
Age (months)		-0.06	-0.06	-0.06	-0.01*	-0.06
Ethnicity (reference group White British)	African	1.83	2.78	2.74	2.72	3.22
	Any Other White Background	0.90	1.21	1.44	1.38	1.50
	Bangladeshi	1.91	2.42	2.66	2.10	2.95
	Caribbean	0.24*	0.87	0.53	0.14*	1.37
	Indian	1.87	2.25	2.57	2.06	2.22
	Pakistani	1.97	2.25	2.06	2.02	2.33
	White and Black Caribbean	-0.24*	-0.24	-0.01*	-0.14*	0.20*
	Other ethnic group	0.88	1.11	0.79	1.01	1.16
Model statistics (including random effects)						
Student level	Sample size	49,319	314,326	46,242	37,181	76,457
	Residual variance	3.59	17.70	6.10	23.70	16.44
School level	Sample size	1,313	3,042	903	732	2,999
	Residual variance	12.75	3.64	24.49	26.88	41.02
Residual school effect (%)		22.0	17.1	19.9	46.9	28.6
Deviance (-2 log-likelihood)		268,384	1,804,081	281,317	230,475	506,384

* = Not significantly different from zero at the 5 per cent level.

Table A1

KS4 course		KS3 science level						
		Missing	2	3	4	5	6	7
Triple award Science	Mean	47.02		23.77	32.16	38.68	44.51	51.52
	N	675		17	206	2,453	15,778	31,950
Dual award Science	Mean	32.67	17.08	21.11	28.59	36.93	43.57	50.93
	N	8,912	481	6,841	35,909	107,109	119,436	51,854
Dual award Applied Science	Mean	26.29	17.96	21.40	28.32	34.93	40.18	46.11
	N	1,924	235	3,024	14,464	23,537	5,724	433
Other Applied Science	Mean	35.19	31.45	32.64	36.00	39.24	42.18	47.32
	N	1,646	172	2,422	11,794	19,009	4,641	406
GCSE Science only	Mean	21.37	15.57	19.96	26.22	33.67	41.08	47.99
	N	6,864	1,150	11,862	29,040	26,408	8,977	1,525

Table A2

KS4 Course	R-square	Predictor	Unstandardized coefficients		Standardized coefficient	t-value	p-value
			B	Standard Error	B		
Triple award Science	0.528	Constant	-11.937	0.257		-46.367	<0.0005
		KS3 fine level	8.701	0.037	0.726	235.805	<0.0005
Dual award Science	0.689	Constant	-8.001	0.058		-138.156	<0.0005
		KS3 fine level	8.113	0.010	0.830	843.333	<0.0005
Dual award Applied Science	0.460	Constant	-2.428	0.176		-13.811	<0.0005
		KS3 fine level	6.408	0.044	0.547	144.101	<0.0005
Other Applied Science	0.137	Constant	20.302	0.232		87.677	<0.0005
		KS3 fine level	3.480	0.045	0.37	78.117	<0.0005
GCSE Science only	0.484	Constant	-6.171	0.132		-46.789	<0.0005
		KS3 fine level	7.354	0.027	0.696	275.406	<0.0005