Value-Added:
from policy to classroom practice

Stephen G. Rogers

Submitted in partial fulfilment of
the requirements for the degree of
Doctor of Education

Supervisor: Professor Gary McCulloch

University of Sheffield

Department of Educational Studies

Doctor of Education Programme (Part 2)

November 2001

ABSTRACT

Using a case-study methodology, this thesis enquires into the development of value-added from policy to classroom practice in a comprehensive school in the North of England. The study is unique in that it examines the work of teachers in relation to policy with a special focus on the extent to which value-added measures can be used to inform an understanding of what is going on at classroom level. It not only links quantitative and qualitative research paradigms, but does so at a level that has received relatively little attention and at a dramatic juncture in the history of teachers' professional lives.

A literature survey shows that although originally conceived as a research tool, value-added was subsequently 'adopted' by secondary schools principally in response to government-imposed 'league tables'. A national value-added scheme has yet to be developed but, in a shift of policy, the government now promotes the use of value-added measures in the new Threshold Assessments of teachers.

Value-added data for core subjects for seven years have been analysed at class level. Pupils in 'top' sets on average obtain positive residuals whilst in 'bottom' sets they are mostly negative. It is shown that this is partly a statistical artefact and therefore not a true reflection of teacher effectiveness. However, when teachers are interviewed they frequently reveal positive attitudes towards upper sets and the opposite with the lower ones.

When value-added scores are considered alongside teacher interview data there are cases where residuals might be indicative of teacher performance but there are no universal patterns. Although there are some indications that pedagogical practice and teachers' backgrounds are linked with the performance of classes, it is concluded that the use of value-added data in the Threshold Assessment of teachers is flawed. Suggestions are made for further research including the use of value-added measures at classroom level.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the backing of, and contributions from, a significant number of people to whom I offer sincere thanks.

Colleagues past and present at Dalestone School have been willing interviewees and have given freely of their time to share their opinions and experiences. Their co-operation and support has enabled me to complete this work at a time when workloads have continued to increase. It has been my aim to represent their views fairly and ethically.

Several members of the research community and colleagues in other schools have offered comments and information at various stages through this research. It was especially valuable to have had the opportunity to attend the 1999 and 2001 BERA conferences. I would like to particularly acknowledge the valuable discussions that I have had with members of the CEM Centre in Durham and John Critchlow, Head of a North Yorkshire School.

I am thankful to the DfES for the award of a Best Practice Research Scholarship in the final year of this research. The financial support at this stage was particularly valuable.

Special thanks go to Gary McCulloch. His sensitivity and encouragement have been unfailing. He has guided me through this research with patience, skill and professionalism. He has given freely of his time and attention to every detail.

The understanding, prayers and patience of many friends have been immeasurable and have been greatly appreciated. Finally, the constant encouragement of my family has been truly magnificent.

CONTENTS

Figures

Tables

Abbreviations

ALIS	A-Level Information System
BERA	British Educational Research Association
CAT	Cognitive Ability Test
CCR	Central Curriculum Record
CEM	Curriculum, Evaluation and Management Centre (University of Durham)
DES	Department for Education and Science
DfE	Department for Education
DfEE	Department for Education and Employment
DfES	Department for Education and Skills
GCSE	General Certificate of Secondary Education
KS3/4/5	Key Stage 3/4/5
LEA	Local Education Authority
NAHT	National Association of Headteachers
NFER	National Foundation for Educational Research
OFSTED	Office for Standards in Education
PANDA	Performance and Assessment (report)
QCA	Qualifications and Curriculum Authority
QUASE	Quantitative Analysis for Self-Evaluation
SCAA	School Curriculum and Assessment Authority
SEN	Special Educational Needs
SHA	Secondary Headteachers' Association
SMT	Senior Management Team
TTA	Teacher Training Agency
TVEI	Technical and Vocational Education Initiative
YELLIS	Year 11 Indicator System

Chapter 1 Introduction

> *...Teaching is the core technology of what teachers do. It is more and more prescribed as politicians and others start to, quite rightly, intervene in the teaching methods that are used. It is the distinct area of teaching that provides for me the most likely explanation for why educational reforms in Britain have hitherto always failed, namely that in pulling the 'lever' of the school we have missed pulling the 'lever' of the teacher. (Reynolds 1998b).*

1.1 Background and Justification

This study is set in the context of the radical modernisation of the teaching profession that was heralded by the Green Paper *teachers: meeting the challenge of change* (DfEE 1998). My focus is the concept of 'value-added' - the research and development of the idea, the development of value-added policy and its application to practice, and a critical engagement with the responses of teachers. Value-added has a relatively short yet dynamic history and is now a key aspect of government policy and probably most LEA and school policies.

This educational case study has two research dimensions linked to a theme of teacher effectiveness. Value-added measurements are a crucial part of the evidence that 'effective' teachers must provide in order to progress onto a new 'upper' pay spine yet Reynolds says that 'our ignorance of the area of teacher effectiveness is virtually total' (1998a:26). The question has been raised as to whether value-added measures *can* be attributed to the effects of teaching or school organisation (Cutler and Waine 1999:60) or indeed as to what value-added indicators really mean (Sparkes 1999:33). Therefore, the first aim of this

research is to address that gap in our knowledge by asking to what extent value-added measures are of worth in the assessment of teacher effectiveness. Further to this, to what extent can we be confident that a particular teacher makes a difference to the educational performance of pupils as measured in terms of value-added?

This study secondly investigates the extent to which value-added policy has had an effect on teachers' attitudes and behaviours; are their 'levers' being pulled or likely to be pulled? Is there a reciprocal relationship – that is, can teachers influence value-added policy or implementation? Such relationships have previously been identified in respect of the National Curriculum and the Technical and Vocational Education Initiative. Osborn *et al* (1997:52-53, citing the work of Croll *et al*, 1994), describe this as 'creative mediation' where:

> ...[teachers] have the ability to transform themselves into 'policy-makers in practice' in the classroom...In this view, teachers have the ability to mediate educational policy in the light of their own beliefs about teaching and the constraints which operate on them in the classroom.

A case study is a powerful tool with which to examine the extent to which Osborn's ideas apply to the concept of value-added. Faced with both internal and external demands to raise achievement, set targets and add value, what choices do teachers make about the way in which they carry out their work? To what extent do teachers 'act communally to make parallel policy to that intended by government' (Osborn *et al.* 1997:53) and filter policy through their own values before it is translated into classroom practice?

Viewed in these terms, this research study raises fundamental problems about the role of teachers, the nature of their professionalism and the measurement of their effectiveness. Reynolds promoted a debate about teacher effectiveness in his 1998 Teacher Training Agency (TTA) lecture only months before the Green Paper was published. Central to his case was an assertion that there should be an applied science of teaching and a belief that good teaching could be measured (Reynolds 1999a). How well does this capture the ways in which teachers respond to changes and exert their own pressure and guidance in the classroom?

Value-added in context

> *'No case study is worth a candle without an understanding of how the educational landscape fits together.' (Gubb 1999:18)*

There are three main components of the value-added 'landscape'. The dynamics of **education policy** have been responsible for its rise from its beginnings in research to its current political, institutional and commercial importance. Secondly, there is the question about whether value-added measurements do, or should, take into account the impact of **the social dimension**. Thirdly, there is the impact of value-added **in the classroom** especially since value-added measurements are now used for target-setting and in Performance Management.

Education policy

In the past 25 years this country has moved from having a largely independent and professionally based school system to one that might arguably be categorised as amongst the most controlled and centralised in the world. In particular, the 1988 Education Reform Act (ERA) created an authoritarian thrust that has been intensified under New Labour. It has been described as 'Bonapartist' by Maden (2000:24), accused of 'stifling innovation' by members of the Institute of Public Policy Research (Slater 1999) and charged with using a 'command and control model' by the Opposition (Willetts 1999).

The former Secretary of State for Education and Employment (1997-2001), David Blunkett, took a firmer grip than his predecessors on some aspects of how education is delivered in the classroom. New Labour policy states that:

> We need to improve the combination of pressure and support which central and local government apply to schools to stimulate constant improvement and tackle under-performance (New Labour National Policy Forum Report, 1999:149.)

Estelle Morris, then School Standards Minister (now Secretary of State for Education and Skills), said that New Labour believes 'in intervention in inverse proportion to success' (Morris 1999:5) - light-touch inspections will only be for those who are meeting government targets.

Reynolds' language was provocative. His use of the word 'technology' in relation to teaching is consistent with his notion of an applied science of teaching. He felt that it is right that politicians intervene in matters of pedagogy. Pejorative statements such as 'pulling the "lever" of the teacher' or 'it's the classroom,

stupid' (Reynolds 1999b) clearly illustrate his message. Inevitably, he has invited criticism. Edwards (1998:30) felt that Reynolds' 'essentially narrow and linear model...trivialises the complexity both of the means and the ends of effective learning' and Lawlor (1999:13) believed that Reynolds' ideas attacked the very notion of teacher professionalism:

> There is no general skills model to fit every circumstance. If there were, then teaching would not be a profession; it would be little more than the activity of a clerk, performing tasks to order.

Hammersley (1999) advocated a more moderate perspective. He expressed concern that the extreme views expressed by Reynolds and Lawlor had such influence on Government education policy. However, it can be argued that what Reynolds attacked was the remnant of old teacher professionalism that predates the Education Reform Act. The introduction of a National Curriculum effectively removed teachers' control of the curriculum, to the extent that this ever existed. Since its election in 1997 and subsequent re-election in 2001, the New Labour government has been moving into the area of pedagogy with, for example, the introduction of Literacy and Numeracy hours. New Labour is introducing a 'modernised' professionalism, very different from that of the past which makes teachers responsible for improving standards for pupils, including those in socially disadvantaged areas (McCulloch, Helsby et al. 2000:1)

Reynolds expressed concern about social inequality in a powerful article in the *Times Educational Supplement* in March 1997. He acknowledged some of the challenges facing educational policy and practice and in particular the issue about how we measure school quality. Reynolds (1997) and the former Chief Inspector of Schools, Chris Woodhead (1998:13) believed that use of 'raw'

school results acknowledges the impact of social disadvantage whereas statistical manipulations of pupil and examination data may 'recalculate absolute failure to be relative success' (Reynolds 1997). Their answer to this issue was to advocate the use of value-added techniques but only in combination with the setting of high targets and retaining the use of actual test and examination results in the assessment of performance.

The publication of 'raw' school examination results in the early 1980s, followed by the introduction of so-called 'League Tables' in 1988, reinforced the idea that pupil performance can be measured by examination results. 'League tables' gave the message that good results are obtained by good schools and *vice versa* - much to the chagrin of many schools and their teachers. Hence statistical techniques such as 'value-added' have been seen by many as a 'fairer way of comparing schools' examination results' (Gray, Jesson et al. 1986:91). In the measurement of school effectiveness, value-added approaches have frequently been shown to be more informative than 'raw' examination results (for example see Fitz-Gibbon 1996:119).

Value-added began as a research tool, and by the mid-1990s had been embraced by increasing numbers of schools and local education authorities (LEAs) and finally politicians. Saunders and Rudd (1999:1) recognise that each 'stakeholder' and user has different expectations and requirements, and crucially that the 'academic debate on school effectiveness and how to measure it [was] now integrally linked with the national political agenda for educational quality'. The Green Paper proposed a new career structure incorporating

rewards for good teachers and new arrangements for appraisal and assessment. Pay that is linked to pupil outcomes was the most controversial aspect of the government's proposals. Some teachers felt that the government was imposing an inappropriate model on education as a whole (Saunders and Rudd 1999). Yet as early as 1995, Jesson (1995:247) speculated that value-added evaluations could be used as a means of identifying 'good practice' and 'might even give some evidence of the effectiveness of teaching of individual teachers'. David Blunkett has said that 'a new system of appraisal would give bigger [salary] rises to those who had brought about the most significant improvements in their pupils' performance' (Sylvester 1998).

Both the Conservative and Labour governments have encountered problems with the value-added concept, mainly on methodological grounds (for example see Hackett 1999c). However, schools have been generally enthusiastic. This has encouraged commercial developments with schools buying statistical services from two main organisations; the Curriculum, Evaluation and Management (CEM) Centre (University of Durham) and the National Foundation for Educational Research (NFER). Such services are available from Key Stage 1 through to Key Stage 5. The government has yet to develop an acceptable methodology for a value-added statistic within performance tables. However, the government is producing data that contribute to the value-added 'theme', for example 'PANDA' reports and the 'Autumn Package'. The PANDA (Performance and Assessment) report is a management tool consisting of benchmark data produced for schools by OFSTED (Office for Standards in Education). Although none of the data is truly value-added, the various indicators focus on 'how well is

my school performing'. The Autumn Package consists of national performance data, summary results, benchmark and value-added information.

Some schools already make extensive use of value-added measures. Haigh (1999:31) reported on a Stafford school whose Headteacher commented that value-added data had 'forced staff to ask questions they've never asked before - about the quiet pupil at the back of the class that we've missed, for example'. 'It makes it possible to make better judgements'. This Head noted that one of the challenging effects of using value-added measures was that they highlight differences in performance between individual departments or teachers. He was aware of the need for Heads to run their schools so that such comparisons are not threatening. He established a climate of trust where no blame was allocated. His assumption was that where the figures reveal a problem, the department, or the individual would themselves see what needs to be done.

However, Bassey (1999b:20), Mortimore (1999) and many headteachers (for example Claydon 1999) believed that the government's proposals would damage that sort of collegial management practice. According to Bassey, the focus on past performance and an overemphasis on quantifiable objectives ignores future potential and human factors in management. Bassey argued that government plans for performance-related pay were based on an outdated model and the CBI (Rafferty 1999) warned that any links with pay should wait until a new Performance Management system has 'bedded down'. The teachers' unions have been somewhat self-contradictory about performance-related pay,

although particularly during the Easter Conferences of 1999 and 2000 they were supportive of members wishing to cross the threshold by, for example publishing guidance for applicants on their websites. The National Union of Teachers obtained a High Court ruling in July 2000 which delayed further progress on the Threshold process until late in the following autumn.

Hence, in order to qualify for a £2,000 increase in salary, teachers must meet standards in five areas of professional practice. One standard, that of pupil progress, has received the most attention from government, teachers and professional associations. Jesson's prophecy has been fulfilled since value-added data will be used in the assessment of whether or not a teacher has met the standard:

> Teachers should demonstrate that as a result of their teaching their pupils achieve well relative to the pupils' prior attainment, making progress as good as or better than similar pupils nationally (DfEE 2000:4).

The Social Dimension

The significance of socio-economic background is the subject of many debates. Widely varying statistics appear depending upon the methodology employed to gather and analyse the data. Using a post-code census approach, Gibson and Asthana (1998b:276) found that over 75% of the variation among schools in the proportion of pupils getting five or more A*-C grades at GCSE could be attributed to social background. At the other end of the scale Thomas and Mortimore (1996:5) found that the significance of socio-economic factors was 'slight'. Using free school meals as the indicator of social deprivation usually gives a stronger political edge to a particular statistic and reports consistently

show an inverse relationship between eligibility for free meals and achievement and a North-South divide (Dean and Thornton 1999; Hackett 1999a; 1999b).

It is in the light of such data that school effectiveness research has been criticised since it explicitly seems to avoid engaging with the socio-economic dimensions of school-level performance. Value added methodology was criticised by Gibson and Asthana (1998a:207) as obscuring the extent to which 'profound differences in the educational outcomes achieved by schools are underpinned by underlying variations in the cultural capital upon which individuals and communities draw'. Reynolds (1997) emphasised this point by commenting that value-added made it possible for low levels of achievement to be regarded as a relatively successful outcome for a school. He continued with a note of realism:

> The problem with the value-added approach, as it has moved away from being a simple research tool to identify effective schools towards being an agency of national education policy, is that it is pupils' actual uncontexted *(sic)* levels of achievement with which they enter our labour markets.

Gibson and Asthana (1998a:195) noted that the government White Paper *Excellence in Schools* (DfEE, 1997) rekindled interest in the contextual factors that influence the quality and performance of schools. For the previous 20 years the school effectiveness movement had been growing and, according to Gibson and Asthana, effectively '..."designing out" a concern with the underlying socio-economic explanations of variations in performance' (1998a:199). Fitz-Gibbon, however, clearly stated that prior achievement measures already include the effects of home background (1996:149) and would dispute Gibson and Asthana's assertion.

Consideration of the social dimension highlights some of the problems that are generated when statistics are used - how are they are calculated and how should they be interpreted? There is much scope for confusion and in the context of Performance Management a need for clarity. Consider, for example, the review meeting (DfEE 1999a) when the teacher of a bottom set of children from a poor part of the school's catchment does not meet his or her objectives. If value-added data have been used, to what extent does this take into account socio-economic backgrounds and the multitude of other factors that might affect attainment such as family breakdowns, illnesses, the timetable, etc? Fitzgibbon (1996:150) emphasised the need to be as fair as possible and that 'value-added measures should be based on cognitive measures and student level data, not on home-background measures and not on aggregated data'.

A focus at classroom level

Schools are progressively being recognised as complex organisations. Ball's (1997a) data from one 'good school' revealed internal differences in values and management that pervaded the operation of the institution. Subject departments are increasingly a focus of interest and are being recognised as being crucial in the quest to raise standards. Although Weick recognised in 1976 that schools were 'loosely-coupled systems' where each unit 'preserves its own identity' (1976:3) interest has been rekindled recently and the importance of subject area culture has been promoted by researchers such as Busher and Harris (2000),

Donnelly and Jenkins (2001, with respect to Science), and Rogers (1997, with respect to Modern Languages).

Just as the direct effects of recent education reforms have been mainly at school level so has much of the use made of value-added data (Saunders and Rudd 1999). With the exception of the Literacy and Numeracy initiatives, neither of these developments has had significant immediate impact on classroom practice. The growing consensus is that it is at this level that attention should be focused. Goldstein (1998:521) has stated that 'as the research progresses, people are beginning to discover that much of the interesting variation occurs at class and teacher level...and also that schools are differentially effective for different kinds of students'. Coe and Fitz-Gibbon (1998:429) found that the contribution to pupil-level variance that is associated with teachers is as high as 42%. This was compared to 'school effects' of between 10 and 20% which led them to conclude that it is the 'teacher that counts; not the school'.

According to Scheerens (1992:8) the main focus of school effectiveness research is the study of pupil attainment levels and the 'predictors' or 'determinants' of these results ('outputs'). These determinants are defined as characteristics of the school. Scheerens noted that these characteristics are the sum total of education in the various classes - 'classroom characteristics are aggregated to school characteristics'. Research by Mortimore and others (1998:94) has considered students' academic results over time; the possibility of differential effects within the same school; and the effects of context on school performance. They found that examination value-added scores are relatively

stable over time but this stability conceals a considerable amount of change within various subjects - what they have termed the 'swings and roundabout effect'. They have uncovered a complex picture of differential effects. Such differential effects occur for students of different prior attainments, different ethnic backgrounds and, to a lesser extent, different gender and social backgrounds.

Thus recent research has drawn attention to the importance of the classroom level and the need to tease out the myriad influences that best predict student progress (see, for example Hill and Rowe 1996; Mortimore 1998:95). This work will build largely on literature from quantitative research and in particular the school effectiveness and school improvement movements. It aims to build upon the work of researchers such as Fitz-Gibbon, Goldstein, Gray, Jesson and Reynolds.

Saunders and Rudd's (1999) approach to their research has similarities to this work. Their case study approach into schools' use of value-added data used information from the NFER's QUASE (Quantitative Analysis for Self-Evaluation) value-added model. The data were derived from questionnaires, published research and a case-study, and the first-hand experience of the authors through their work with schools. Their focus was on how schools perceived and used QUASE data down to departmental level. This included an exploration of policy issues and the role of intra-organisational factors. Their study did not, however, enter into the territory of the results of individual teachers. Saunders and Rudd

reached conclusions that concentrated on the use of data for management purposes whereas this study has a focus on teacher effectiveness.

The government recently spent (it is reputed) between £3 and £4 million on a study of teacher effectiveness by management consultants Hay McBer (Barnard 2000a:21). Using in-depth interviews and observation techniques and matching the findings with pupil progress, Hay McBer identified 16 attributes that contribute to effective teaching (2000). However, there were no surprises - the attributes can all be found in earlier research findings. Of greater significance however, is the response of a special seminar of the British Educational Research Association (BERA) held in May, 2001 where serious methodological shortcomings in the Hay McBer research were highlighted (BERA 2001). One of the conclusions from the seminar was that the Hay McBer research gave only 'limited insight into the complex and multi-varied work of teachers' (ibid. p.9).

The present study is unique in that it examines the work of teachers in one school in relation to policy with a special focus on the extent to which value-added measures can be used to inform an understanding of what is going on at classroom level. It not only links quantitative and qualitative research paradigms, but does so at a level that has received relatively little attention and at a dramatic juncture in the history of teachers' professional lives.

1.2 Methodology

For this research an educational case study approach has been adopted because of its 'capacity for understanding complexity in particular contexts'

(Simons 1996:225). The case study was of the 'instrumental' (Stake 1998:88) or 'theory-seeking' (Bassey 1999a:62) type - intended to provide insight into particular issues. This approach has been chosen because it is 'strong in reality'; because case studies recognise the complexity and 'embeddedness' of social truths; and because such a study can contribute both to the school and to educational policy-making (Cohen and Manion 1994:123). This method has also been chosen as the best balance between practical concerns relating to time and access and the research focus. Being strong in reality and appreciating the complexities of the classroom situation are important features of this research. Hence the voices of teachers are important:

> They have access to the rich contextualising information that might make some sense of the inherently variable outcomes, pupil by pupil, syllabus by syllabus (Coe and Fitz-Gibbon 1998:434).

It is expected that this work will contribute to the debate about teacher effectiveness and that any theory which is developed can be 'grounded' in hands-on experience with practical school improvement efforts (Scheerens 1992:68). An aim is for this study to reveal 'processes at work that may or may not be present in all high schools but which are likely to be present in many of them' (Arksey and Knight 1999:58).

The research is also strongly influenced by my position as a Deputy Headteacher and as the parent of two children at the large comprehensive school in which this study is based. Like Gubb (1999:18) I see case study as building a dense and textured picture of the complex worlds within a classroom. However, this research did not include direct classroom observations partly as these tend to damage the 'delicate web of classroom interactions' (Gubb

1999:18) and partly to limit the scope of the study. Rather, this research focuses on teachers, their reflections and the value-added outcomes of the pupils that they teach.

A methodology has been developed to explore the confluence of a number of areas that have been highlighted in the literature by researchers including Harris, Jamieson and Russ, Saunders and Rudd, Reynolds, Coe and Fitz-Gibbon, Thomas and Mortimore, and Ireson and Hallam, as requiring study.

Harris *et al* (1995:297) recognised the need for studies to confirm whether their findings about effective departments could be replicated. They also called for 'finer-grained' studies 'so that we can see whether there are distinctive features of different subject departments.' Harris *et al* also recognised that smaller scale studies that operate at departmental or classroom levels were closer to the core function of teaching and learning and complemented large-scale, whole school research.

Saunders (2000) and Saunders and Rudd (1999:4) believed that the enormous national investment in performance data has been something of an act of faith. They suggested that research is needed into how and for what purposes teachers are using value-added data. They added that an 'understanding of the "'micro-political" context may be highly relevant' (ibid. p.10).

Reynolds (1998a) mentioned the 'urgent' need for research in contemporary schools. He (1999b) underlines the importance of research at the teacher level.

His logic was that we know that the teacher 'level' explains three to four times more in terms of pupil results than the school 'level'. We also know that there is 'substantial variability in quality within our schools, in terms of departmental performance and in teachers' effectiveness'. However, he believed that we have largely ignored the teacher because Britons historically have difficulty in addressing the interpersonal and political issues related to variation in the competence of teachers. Reynolds further believed that we need to analyse the variation between teachers and use this variation to identify effective ways of teaching. He suggested that it is teacher behaviour that matters more than the organisation of classrooms.

Coe and Fitz-Gibbon (1998:433) called for better evidence about how and how much schools and teachers can influence the outcomes that we are measuring. They recommended that research should focus on forming much better theoretical understandings of how different features of students, classrooms, schools and their contexts may cause different effects. The lack of longitudinal data is something that they and Sammons (1999:27) lament. Thomas and Mortimore (1996:28) specifically mention the need to investigate and describe the relationship between negative and positive value-added results and measures of school processes, in particular the quality of teaching and learning.

There is also a need to build on recent research such as that by Ireson and Hallam and Boaler into the effects of ability grouping. Ireson and Hallam (1999:354), for example, expressed the need for a 'clearer picture of the relative effects of grouping on both academic and non-academic outcomes for pupils'.

They also called for 'a better understanding of the way in which grouping is related to the ethos of the school, to teacher and pupil attitudes and to classroom teaching (ibid. p.354).

This research will highlight the changing nature of teacher professionalism and in particular the impact of increased accountability. As efforts are made to measure teacher effectiveness, the present study demonstrates the influence, utility and value of using value-added data in this pursuit.

1.3 Definitions

Fitz-Gibbon (1996:13) explained that the term **'value-added'** is the 'fashionable' way in which we in the UK speak about the statistical term 'residual' which:

> ...is defined as the difference between the result obtained and the result predicted from measurements of factors known to be correlated with the outcomes.

A negative residual indicates a worse than predicted performance. Raw residuals are usually rescaled (standardised) to show their position in a whole sample. Fitz-Gibbon used a more widely understood definition in the Value-Added National Project Final Report:

> Value-added was defined for each pupil as the difference between a statistically-predicted performance (based on prior attainment and the general pattern in the data) and the actual performance (Fitz-Gibbon 1997:3).

Fitz-Gibbon (1996:119) warned about misuse of the term 'value-added'. Although she accepted the term value-added 'with reservation', she preferred to use 'residual' since it is a statistical term and does not pre-empt an interpretation. In this research however, the term 'value-added' is preferred

since it is the one which is embedded in public and political debate and to which most teachers can relate.

In keeping with Fitz-Gibbon and the former School Curriculum and Assessment Authority (SCAA), the starting point for any analysis in this research will be the data for individual students. This is because indicators of value-added look at the *progress* (original emphasis) made by pupils over a given period (SCAA 1997:3). In this way the value-added indicators compare the progress made by a pupil with the progress actually made by pupils generally in a given sample (ibid. p.3). The scores are thus measures of relative progress, *comparative* or *relative* value-added. These terms are a more accurate description of residuals but are not in common usage and will not be used in this thesis. However, it is important to note that the sample with which the school measures are compared is a national one. In this study these are relative value-added data and recall of this will help the reader appreciate the strength of the arguments presented.

Value-added measures can be averaged to give indicators of class, departmental or whole school value-added. It is in this context that value-added research has played a key part in the **school improvement** and **school effectiveness** movements. Hopkins *et al.* (1994:3) emphasised that 'school improvement is about raising student achievement through focusing on the teaching-learning process and the conditions that support it'. In contrast, their definition of school effectiveness was in terms of:

> ...the differences in student outcomes... that schools achieve after full account has been taken of the pupil's prior learning history and family background at the time he or she enters the school. In terms of the contemporary debate over league tables, this is the 'value-added' to a pupil over and above what ability and socio-economic status would naturally bring him or her (ibid, p.44).

It is not insignificant in the context of this study that researchers have found that a synthesis of these very different paradigms has created some intellectually creative and practically productive studies. Reynolds *et al* (1993:51) called for rich case study data to improve the practitioner relevance of the effectiveness research and to facilitate the transfer of knowledge to the improvement community.

Although Ozga (2000:2) argued that there are different ways of defining **'policy',** her broader view of it being a *process* ('involving negotiation, contestation or struggle') rather than a product will be adopted in this research. This view permits the actions of different groups, including teachers, outside of the 'formal machinery' to be considered as contributors to policy-making.

I define **'practice'** for this study as what goes on in the classroom - the application of pedagogy. The definition by Creemers (1994:10) of 'instruction' to mean 'education at classroom level in a broad sense' is a close approximation to 'practice'. Creemers included teacher-initiated activities and, most importantly, teacher behaviour but admitted that the term instruction is not widely accepted and that it is often used in a narrower sense. I include the planning, teaching and class management strategies employed by teachers; aspects that are

regarded by the Teacher Training Agency as important since they feature in the Standards for Qualified Teacher Status (TTA 1998).

1.4 Delimitation of scope

The conclusions from this study are expected to contribute to the current debate about teacher effectiveness and the professional identity of teachers. With its focus on the issue of value-added, it is anticipated that this research will be of value as 'Performance Management' is institutionalised. This work will therefore concentrate on teachers, their classroom practice and the resultant pupil outcomes in value-added terms. Other aspects of teachers' lives, for example management responsibilities, will not be considered unless they have a direct impact on the research. Thus the policy and practice of placing of pupils in teaching sets would be considered relevant.

In their study of students' experiences of ability grouping, Boaler *et al.* (2000) were aware of the fact that they had not, at that time, interviewed teachers. They were sensitive to the potential criticism that their research, based on lesson observations, questionnaires and interviews with students, was 'one-sided' (p.13). In considering data from teachers rather than pupils, this research might be considered to be in some ways complementary to that conducted by Boaler *et al.* In the same way I hope to have captured a reasonably faithful picture of the day-to-day realities of the classrooms, and in particular the teachers, that I am studying.

The exclusion of pupils from the research reflects the practical difficulties of this practitioner research. This study is also restricted to Mathematics, English Language and Science where the data is richest, relates to whole cohorts of pupils and covers a seven-year period – other subjects have been specifically excluded. Reference to English as a subject in this study means English Language. Literature has been excluded since all pupils do not study it.

In order to preserve confidentiality the name of the Local Education Authority and case study school have been changed. Individual teachers will only be identified by a letter, their gender and subject speciality.

1.5 Outline of the thesis

Chapter 2 gives an account of the methodological approach employed in this case study of a single school. It describes the sample and the analysis of a quantitative data set covering seven years, the three core subjects and over 3,850 individual pupil results. The qualitative techniques and sources of other data are discussed. It especially examines my position as Deputy Headteacher of the school and the ethical and validity issues that arise with participant research.

Chapter 3 examines relevant literature to trace the development of value-added from its origins as a research tool, its subsequent adoption by many schools and emergent role in education policy. This chapter includes a critical review of how

value-added has developed into a concept that occupies a central position in school effectiveness policy.

Chapters 4 to 7 form the core of this thesis. In chapter 4 the history of value-added at Dalestone School is briefly traced. This chapter explores the way in which the use of value-added data has developed at the school in relation to wider changes in policy and draws substantially from documentary and anecdotal evidence. It includes a consideration of the roles of certain staff and governors and recent developments in respect of target-setting.

The qualitative data set forms the focus of chapter 5. The value-added data for Key Stage 4 English, Mathematics and Science sets are considered and compared both on a longitudinal basis and across the subjects.

Data from teacher interviews are discussed in chapter 6. These are considered in relation to teaching sets and across subjects. Teacher views on pedagogy and their views on professional matters form a particular consideration of this chapter.

Chapter 7 brings the different data sets together in order to explore the extent to which value-added measures may be of worth in the assessment of teacher effectiveness and the effects of teacher attitudes and behaviour on the educational performance of pupils as measured in terms of value-added.

The final chapter presents the conclusions of this study and examines the extent to which the research questions have been answered. The ways in which value-added policy has had an effect on the work of teachers and the reciprocal effects of teachers on policy are examined. This chapter assesses the implications for theory, policy and practice and especially considers the issues associated with performance related pay for teachers. Recommendations for future research are made.

Chapter 2 Methodology

> *We need to know more about the factors associated with success and failure of both students and institutions, but this is a painstaking, long term and complex process. Unfortunately, we appear to be passing through a phase of our culture where those in authority, or who wish to be in authority, have little taste for confronting the complexities of the real world in favour of oversimple interpretations. If such interpretations are not challenged, they may distort and degrade the systems they are supposed to support and describe (Goldstein 1997c:19).*

2.1 Introduction – research questions

The main methodological device used in this thesis is that of a case study. This chapter explores that choice and its appropriateness to the chosen research field. This is a case study with two key methodological strands - an analysis of value-added data, which is quantitative in nature, coupled with semi-structured interviews of teachers which are qualitative.

The previous chapter referred to the growth of value-added in the measurement of school effectiveness. The context for this research is problematical with a wide range of views about the value and utility of value-added data existing. The view that value-added approaches are more informative than 'raw' examination results (Fitz-Gibbon 1996:119) is particularly widely held amongst researchers and schools. However, politicians generally remain sceptical and several in the research community (for example Goldstein 1997c) warn about the statistical difficulties associated with value-added. Jesson (1995:247) speculated that value-added evaluations could be used as a means of identifying 'good practice' and 'might even give some evidence of the effectiveness of teaching of

individual teachers', but it is only in the area of Threshold Assessment that government scepticism has significantly dissolved. In their commentary about problems with respect to the use of examination results, Cutler and Waine (1999:60) noted that problems remain - not least they question whether value-added measures can be attributed to the effects of teaching or school organisation.

It is in the disaggregation of value-added data that many 'problems' arise. Examination value-added scores for a school tend to be relatively stable over time (Mortimore 1998:94), but single statistics are of limited value and mask the variation that may occur within various subjects. Sammons *et al* (1997) for example noted significant differences when data are analysed at department level. The question of whether schools are equally effective for all pupil groups has been raised. Researchers such as Schagen and Morrison (1999:8) noted a 'common picture' of 'higher ability students overachieving while less able students attained lower grades than expected', and suggested that this might be explained in terms of different pedagogical treatment of students in different sets.

There has been relatively little work using 'disaggregated' value-added data. One of the most extensive studies was that by Sammons *et al* (1997) in which data were analysed using multilevel techniques at the level of individual departments. However, this did not go down to the level of individual classes.

It is particularly interesting to note that the government is happy for value-added data to be included in Threshold applications (albeit only one aspect of the assessment process) yet there is a dearth of research evidence to show whether or not the classes of good teachers obtain positive value-added scores. This is one crucial area where relevant research is needed. Others were outlined in chapter 1 and, briefly, they include:

- ◊ A need for smaller scale, 'fine-grained studies' at classroom/teacher level
- ◊ How and for what purposes are teachers using value-added data?
- ◊ A need for longitudinal studies
- ◊ A need for a clearer picture of the effects of ability grouping

This research aimed to explore the territory between teachers and the value-added scores that pupils in their classes obtain. It asked:

- ♦ To what extent can value-added measures be of worth in the assessment of teacher effectiveness?
- ♦ What are the effects of teacher attitudes and behaviour on the educational performance of pupils as measured in terms of value-added?
- ♦ What is the relationship between value-added policy and teacher attitudes and behaviour?

2.2 An Educational Case Study

For this research an educational case study approach was adopted because of its 'capacity for understanding complexity in particular contexts' (Simons 1996:225). The case study has been of the 'instrumental' (Stake 1998:88) or 'theory-seeking' (Bassey 1999a:62) type - intended to provide insight into particular issues. This approach was been chosen because it is 'strong in reality'; because case studies recognise the complexity and 'embeddedness' of social truths; and because such a study can contribute both to the school and to educational policy-making (Cohen and Manion 1994:123). This method was

also chosen as the best balance between practical concerns relating to time and access and the research focus.

Perhaps of greatest significance for this thesis is the uniqueness of case study as a form of research to focus in depth and from a holistic perspective yet generate both unique and universal understandings - what Simons calls the 'paradox' of case study (1996:225). Arksey and Knight (1999:58) also endorsed the possibility of generalisation from case study. They observed that it can, for example, 'show processes at work that may or may not be present in all high schools but which are likely to be present in many of them'.

Case study has also been appropriate because it has no specific methods of data collection and is not governed by traditional views of data collection (Bassey 1999a:69). Thus this study has drawn upon a variety of data-collecting methods. These are:

- ◊ The collection of value-added data by subject, set and teacher - a quantitative technique
- ◊ Teacher interviews - a qualitative technique
- ◊ Observation and documentary evidence for the purpose of triangulation. This specifically included evidence resulting from my position as Deputy Headteacher of Dalestone School.

A flow diagram summarising the research design is shown in Figure 1.

2.3 Dalestone School - the sample

Dalestone is a co-educational, comprehensive school with around 1220 pupils on roll and a full-time equivalent of over 70 teaching staff. The school serves a


2a
2b

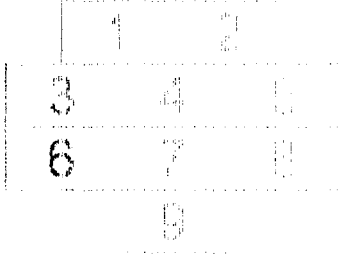
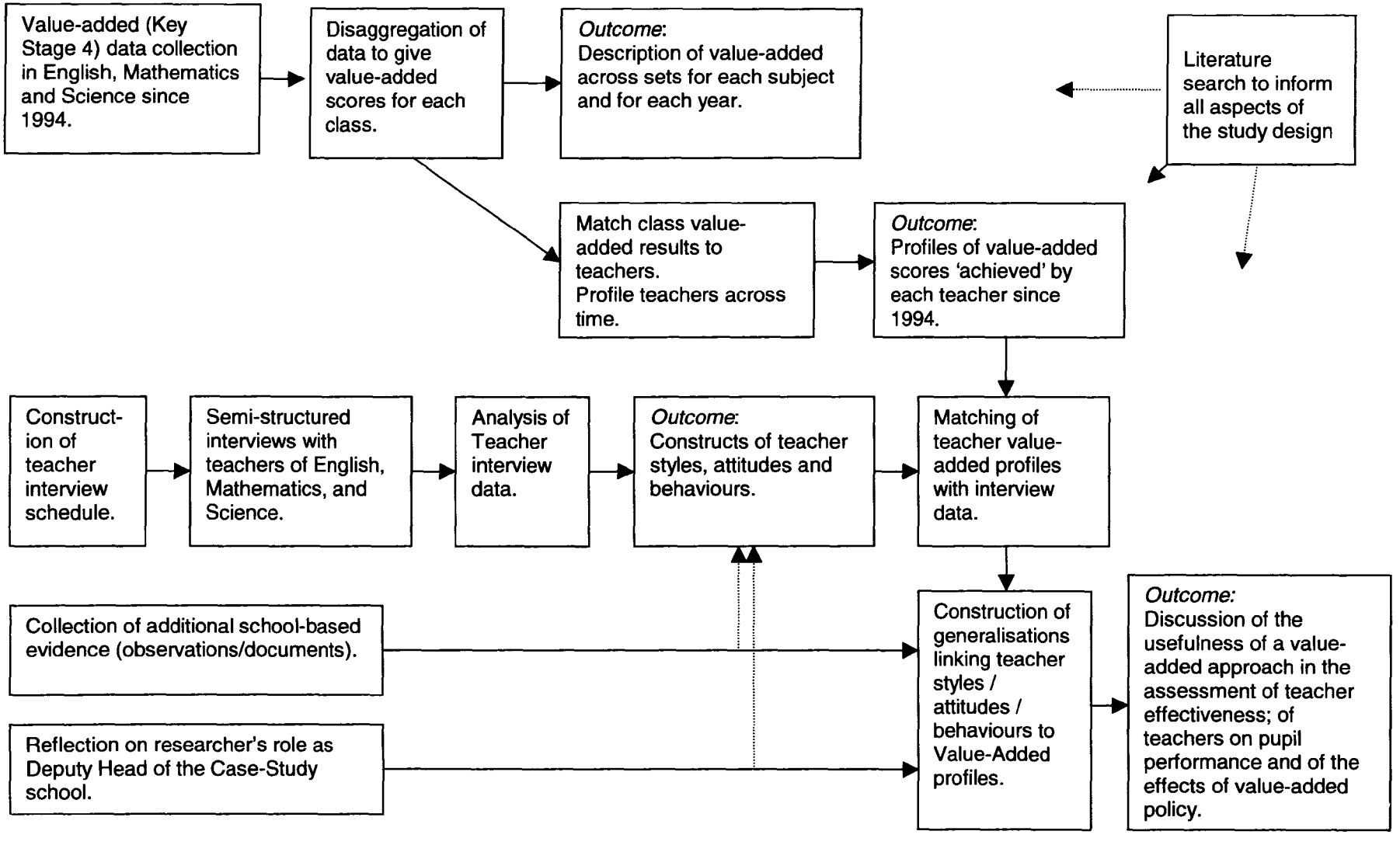

1
2
3
4
6
7
9

Figure 1 Outline of Case Study data collection process



part-rural, part-commuter, socially mixed community covering approximately 120 square miles. The school was chosen for this case study because it is not complicated by factors such as selection and competition. A local independent school and other neighbouring schools attract very few from the potential pupil intake; pupils mainly come from 11 relatively small partner primary schools but there is very little exercising of parental choice.

Dalestone is an all ability school with a 'mission statement' that promotes 'equal worth' and 'achievement for all'. As such it accords with the definition of a comprehensive school as given by John Prescott MP, Deputy Leader of the Labour Party (1997:198). It is also comprehensive in that it essentially educates all of the local children (Walford 1997:55). In keeping with the trends recorded by Gewirtz *et al* (1995:169), Dalestone pupils have been increasingly taught in ability sets rather than the mixed ability groups which are seen by some as part of the comprehensive ideal. Ball (1997b:72, writing when the Conservatives were in power) argued that setting was a response to policies of the market and choice in education. Writing about New Labour's education policy, Power and Whitty (1999:540), however, saw it as an acceptance of inequality as a 'fact of life'. Hence Dalestone can be seen as a mixture of the different basic philosophies of comprehensive education suggested by Ball (1981:6-10) that Burgess (1983:239) considers being typical of most comprehensive schools.

Less than 0.5% (January 2000) of Dalestone's pupils come from ethnic minority groups. Hence, in common with Abraham's (1995:29) research, neither racial nor ethnic differences will be significant factors. 4% take free school meals and

7.5% are on the Special Educational Needs register. 'Generally attainment on entry is slightly above average, although the number of very able pupils is not great' (Baxendale 1996: paragraph 1.2). Examination results (for example 5 A*-C grades at GCSE) are above both national and LEA averages although comparisons with similar schools (PANDA data) suggest under-achievement in a number of areas. This was a significant aspect of the 2000 OFSTED inspection where a key conclusion was that the school was 'judged to be under-achieving' (Drew 2000:8).

This school has a relatively stable teaching staff although 'turnover' has been greater in recent years. The current headteacher has been in post since Easter 1997; his predecessor had been Head for 23 years. I have been Deputy Headteacher since September 1990 and the study design included my reflections as both the researcher and a senior manager within the school. Further, as a senior manager I have access to documents and other information that might not normally be available to researchers. This is considered to be significant in that it will greatly add to the depth of this research.

The school has been part of the YELLIS (Year 11 Indicator System) project since 1994. Value-added data for the core subjects (English, Mathematics and Science) are available from 1995 and has been analysed at teacher level. These subjects have been chosen since they are studied by all pupils - thus giving a large database. Each cohort of pupils is usually divided into eight or nine ability sets. Thus three subjects over seven years have yielded data for around 174 classes at Key Stage 4 - this includes approximately 3,855

individual subject GCSE results. Twenty-five teachers in the chosen departments were interviewed. This sample size is more than adequate against the guidelines given by Cohen and Manion (1994:89-91). It provides a rich database comparing very favourably with, for example, the major study conducted by Hay McBer. The latter drew pupil progress data from only two curriculum areas and had 96 teachers in its final analysis (BERA 2001:7-8).

2.4 The Research Design

Schagen and Morrison (1999:10) say that:

> ...if school management is serious in developing a programme of school improvement, then it needs to integrate high-quality performance data analysis...with other evidence, including classroom observation... This is not an entirely straightforward process.

They were concerned that consideration of the performance data alone leads to 'sterile arguments' that will actually become obstacles to school improvement. Gibson and Asthana (1998b:269) go further by stating that 'uncontextualised performance statistics are fundamentally flawed'. So value-added analyses need to be considered in the context of the school's background and classroom practice, and in the role that school management, policymakers and others play in the progress that individual pupils make. This methodology aimed to do that.

The methods used in the quantitative (value-added) analyses expected to produce data that could be related to individual classes and hence individual teachers. The teacher interview (qualitative) schedule was structured so as to

utilise the data from the quantitative work. Similarly, the analytical categories were matched to the value-added profiles.

2.4.1 Literature search

A review of relevant literature will be given in chapter 3. In this section particular consideration has been given to research that has informed this methodology.

A relatively well-defined group of researchers have conducted the bulk of value-added research in this country. This includes Fitz-Gibbon, Goldstein, Gray, Jesson and Reynolds. Much research has also been completed on teachers and classroom practice (for example by Peter Woods). In recent years there has been a distinct shift of focus by the research community towards the notion of teacher professionalism and a focus on the life and work of teachers (for example Hargreaves 1994; Helsby 1999; Day, Fernandez et al. 2000; McCulloch, Helsby et al. 2000). This is understandably linked with concerns about the possible erosion of teachers' professional status following the 1988 legislation and more recently the Green Paper *teachers: meeting the challenge of change.*

Most of the value-added studies have been large in scale. Early work by *Gray et al* (1990), for example, involved six Local Education Authorities one of which provided data for over 4,000 pupils. In a comparison of Examination Boards, Tymms and Fitz-Gibbon (1991) drew upon data from their ALIS database covering 70 schools and colleges. This covered 11 advanced-level subjects

over 5 different Examination Boards - again around 4,000 pupils. Yet another study (Gray, Jesson et al. 1995) was based on three successive cohorts of pupils passing through some 30 English secondary schools. Essentially, this was a period when methods of measuring and reporting on value-added were being developed and the main focus of attention was on comparing schools with each other. Jesson and Gray's (1991) study tried to go further but provided no conclusive evidence of differential effectiveness as their work was hampered by the lack of suitable measures of prior attainment.

As finely graded baseline measures were developed (including, for example, the YELLIS test, Cognitive Ability Tests and National Attainment Tests) so more detailed studies were possible. Large studies such as that initially involving 11,881 pupils from 87 schools by Thomas and Mortimore (1996) continued. Subsequently, Thomas *et al* (1997:453) reported that there was increasing interest in the topic of differential effectiveness. Although still large, studies such as theirs still tended to only go as far as the identification of patterns of results for different groups of pupils.

Anecdotal evidence shows that many studies have been completed within schools using data and guidance provided by, for example, YELLIS (Curriculum Evaluation and Management Centre 1999) or using texts such as that produced by Schagen (2000). However, these have been for internal use and consequently not published. The guidance given in the *YELLIS handbook* was used in the interpretation of value-added data in this study.

A paper which has influenced this research design is that by Schagen and Morrison (1999) on judging departmental performance in schools. The main thrust for their work was to compare the results within a given subject area for a school with 'what would have been expected' - 'the latter notion [was] derived from analysis of results for the subject area across a large set of schools' (ibid. p.4). Although they accepted that the recommended technique of analysis for their data set was to use multilevel modelling they rejected this for reasons of presentation (ibid. p.5). The results of multilevel analyses can be difficult to interpret but also it is hard to see what the implications of the model are for different students or groups of students. They consequently sacrificed some statistical rigour for a more graphical display based on a regression model. Although multilevel modelling is a valuable technique, it is likely that the 'loss' has been minimal for two reasons. Fitz-Gibbon (1996:130) firstly points out that there is a variety of statistical reasons why the technique should be treated with caution. For example, a good set of results with a small group of students would be adjusted downwards - this would not be useful in the context of discussing the data with teachers. Secondly, Fitz-Gibbon (ibid. p.131) notes that as the size of a department approaches 30 or more, the difference between a multilevel residual and the simple regression one will be negligible.

For the same reasons this study only made use of relatively simple statistical techniques and graphical presentations of data where appropriate. The potential use of this research by teachers is a further reason for the use of uncomplicated techniques - it would be unreasonable to expect most teachers, governors or politicians to have knowledge of sophisticated statistical techniques. The

government's 'Autumn Package' in which a principle of transparency is apparent - making the data accessible to numerically competent people - illustrates the latter point.

Teacher effectiveness has been considered from a wide variety of methodological positions yet, 'despite the diversity of approaches, there is a degree of consensus about the generic features of effective teaching' (Harris 1998:169). Harris stated that many studies belong to a 'process-product' paradigm (ibid. p.171) where teacher characteristics (process) are correlated with student outcomes (product). The shortcomings of such correlations will be discussed in the next chapter.

In considering this methodology, and in particular the qualitative aspects, the advice of Woods (1996) was of great importance. His promotion of 'symbolic interactionism' as an approach for the exploration of teaching has relevance for this study and accords with Hargreaves' (1996) call for teachers' 'voices' to be heard. Hargreaves (ibid. p.15) nonetheless warns that the priority should be to understand teachers' voices rather than romanticizing them.

Symbolic interactionism was not defined by Woods in a single sentence; perhaps the nearest to a definition was that with symbolic interactionalism,

> The emphasis is upon the construction of meanings and perspectives, the adaptation to circumstances, the management of interests in the ebb and flow of countless interactions containing many ambiguities and conflicts, the strategies devised to promote those interests, and the negotiation with others' interests that is a common feature of all teaching situations (Woods 1996:7).

Woods' explanation (ibid. chapter 2) of the methodological implications of symbolic interactionism demonstrated that it is appropriate for case study and especially since my position in the school has significant implications. For example, Woods discussed the importance of grounding the enquiry in the empirical world under study; reaching behind the 'public façades'; and appreciating the culture and people's innermost feelings, impulses and passions. He also emphasised the researcher's own position and the dilemmas that arise in that role - which Woods believed could be a source of strength if handled sensitively (ibid. p.54).

2.4.2 Value-added data

A key feature of current OFSTED inspections is an assessment of the extent to which a school is self-evaluating. In order to do this, schools need detailed and reliable data. Carol Fitz-Gibbon launched an A-level Information System (ALIS) in 1983 and after a slow start had recruited 348 schools and colleges by 1993. The Year Eleven Information System which started in 1993 with 10 schools boasted 1,223 registered in 2000. Other projects focus on Key Stage 3 and primary schools. The CEM Centre now claims to be the largest provider of performance indicators in the world.

The success of these projects is linked to factors that generally make any enterprise successful - a quality product, service and reliability. YELLIS is concerned with Key Stage 4 (ages 14 - 16, years 10 and 11 of secondary education) and follows a one- or two-year cycle.

Results from a YELLIS assessment test in year 10 (comprising a numerical section and a verbal section) serve as a baseline and correlate strongly with GCSE performance. The test is designed to measure 'ability rather than achievement, fluency rather than knowledge' (Curriculum Evaluation and Management Centre 1999:3). YELLIS predicted grades are generated for each subject by developing a regression equation showing the relationship between YELLIS test scores and achieved GCSE grades in the past.

Schools are given feedback from the test. This includes the placement of each pupil into one of four broad ability bands, predicted GCSE grades for each subject, and 'chances graphs' which indicate the actual range and frequency of grades that were achieved in the past - thus providing a rough guide to what might be reasonably expected. Some staff at Dalestone School make use of this data in the process of target-setting although usually the YELLIS feedback arrives too late for this. Teaching sets are generally formed using data from Key Stage 3 and teacher recommendation. These data are increasingly being used as confidence has grown in the reliability of Key Stage 3 National Test results. YELLIS test scores are frequently used in consideration of any subsequent changes to the composition of sets. Boaler (1997a:578) emphasised that quantitative research will not give an insight into the ways in which setting influences achievement or the processes by which it takes effect. Hence teacher interviews and other information, for example about setting policy in each subject area, will be crucial in this respect.

At the end of Year 11, GCSE results are matched with the Year 10 predictions. Student by student lists show the extent to which each candidate did better or worse than predicted - these 'residuals' are the measure of value-added for each student and an indicator of the net effect that schools have on the progress of a pupil. Further analysis of the YELLIS data provides averages for each subject, relative ratings, and comparison graphs. 'Standardised residuals' are raw residuals that have been adjusted so that they are on a comparable scale with standardised residuals from different subjects or different years. Thus reliable comparisons can be made across the curriculum or over time.

An extension to the basic YELLIS service provides schools with analyses of pupil attitudes and aspirations. More recently, the YELLIS scheme has been piloting the provision of value-added information from Key Stage 3 to Key Stage 4.

In this research the standardised residuals for each cohort in each subject have been sorted into the original GCSE classes. Using archived copies of the school timetable, the teacher(s) for each class were identified. This analysis has provided two main outcomes:

- ◊ A profile of value-added scores for each set in each subject for each year. Thus variations from 'top' to 'bottom' sets can be described, as can variations within and between the three subjects and over time.
- ◊ A profile of the value added scores achieved by the classes taught by each teacher over time.

One of the first issues that this sort of analysis raised was that of statistical significance. With a small number of pupils in some sets can any relationship in the data be inferred or is the distribution purely a chance phenomenon? Fitz-

Gibbon (1996:124) suggested that data pertaining to groups of students should be used to generate hypotheses. Goldstein and Thomas (1995:37) raised the issue of analysing data relating to a small number of students but concluded that as long as the user is aware of any uncertainties then the value-added estimate may be useful as one piece of information among others. Mayston and Jesson (1988:323) reminded us that statistical techniques (in their example regression analysis) in themselves might lack any underlying theoretical justification as adequately modelling the educational process. A correlation thus may not justify a causal interpretation or indicate the direction of any relationship. Thus in a relationship involving positive value-added and students with high baseline assessments, did good teaching directly result in the added value, or did able students stimulate good teaching? Many researchers (for example Thomas, Sammons et al. 1997) stressed the need to examine data over a period of time as fluctuations from one year to the next are to be expected.

The analysis of class data followed the same process as the analyses sent by the CEM Centre to schools. Thus the class standard residuals are presented as a bar chart with two standard deviations being an indication of significance (0.05 level) (Fitz-Gibbon 1996a:126). The formula used to calculate the standard deviations was:

$$\sigma = 2(1/\sqrt{n})$$

where $\sigma = 2$ standard deviations and n = number of pupils in the sample.

Any data that is outside this boundary might be considered to be due to factors other than chance such as good or poor teaching.

Goldstein (1983) and Gray *et al* (1995) have raised issues concerning analyses of performance over time. The following are some issues that were relevant to this study:

- ◊ Each cohort of pupils is different and there may be changes, for example in the social mix of pupils, over time.
- ◊ Educational standards may have changed.
- ◊ Changes may have occurred in the GCSE examinations - for example standards or the Examination Board used.
- ◊ Internal to the school there have been staffing changes and changes to the school day and timetable structure.
- ◊ Changes in educational policy have filtered down to classroom level over time.

Fitz-Gibbon (1996a:126) encouraged drawing from value-added data over time in order to monitor effectiveness. By virtue of the baseline assessment it could be argued that value-added data takes into account, for example, changes in the spread of abilities of pupils over time. Hence the most significant time-related issues are those related to staff changes. However, it is probable that there are many subtle policy and time-related effects. For example since 1997 the school has been responding (initially through training, then through application) to target-setting legislation. This will be having an increasing impact at classroom level as teachers are required to set targets, justify them and monitor progress that pupils make towards them.

Data over time will also illustrate any 'swings and roundabout effect' (Mortimore 1998:94). This has a bearing on the usefulness of this research to policymakers

both within and external to schools in that value-added data is retrospective, hence the classes to which the data relate have finished and the staff concerned may have left the school.

2.4.3 Teacher Interviews

Interviews were chosen as the most appropriate method for revealing teachers' opinions and thinking. Interviews also directly record the 'voices' with which teachers speak, something that Hargreaves (1996:12) amongst others believed has been ignored by policymakers and some researchers. Interviews were of a semi-structured nature and taped. They were conducted between May 1999 and April 2001 - a period following publication of the Green Paper and whilst Threshold Applications were being made and assessed but before any outcomes were declared. They were conducted on an individual basis and covered teacher experience, knowledge and use of value-added data, and issues raised by the Green Paper. A particular focus of the interviews was the actual value-added results of classes taught by the teacher. These were matched to his or her expectations, teaching styles and perceived rapport with each group.

The initial reason for selection of a teacher for interview was the availability of value-added data for at least two classes that he or she had previously taught at the school. This effectively excluded those staff who were relatively new to the school or who did not teach at Key Stage 4. The latter did not actually exclude anyone since all staff in the core subject departments taught at KS4. The

second criterion for selection was that the teacher should be eligible for Threshold Assessment - point 9 or above on the teachers' salary scale at 1st September 1999. One exception was made to this in the case of a Mathematics teacher who was a mature entrant to the profession 5 years ago. None of the teachers who were approached declined to be interviewed.

Five teachers who met the criteria had left the school before interviews took place. They were contacted by letter and completed a questionnaire that closely matched the interview schedule. Four of these teachers were later met on an informal basis when some of their statements were elaborated. Clearly, the responses from questionnaires were not as rich and detailed as face-to-face interviews.

Table 1 shows the distribution of the interviewees and the average length of teaching experience. Note that experience is considered as total teaching experience and not simply years at Dalestone School. The gender balance reflects that existing within the school, which is with a dominance of female teachers in English and males in Science.

Table 1 Distribution of Interviewees.

Note: one male English teacher also teaches Mathematics but is only included in this table under English.

Source	*Gender*	*English*	*Mathematics*	*Science*	*Average experience by gender*
Interview	*Male*	*1*	*3*	*6*	*23*
	Female	*5*	*4*	*1*	*18*
Questionnaire	*Male*	*1*	*0*	*1**	*29*
	Female	*0*	*1**	*2**	*24*
Average experience by subject		*19*	*24*	*22*	
Total		*7*	*8*	*10*	

** = those subsequently met informally*

The advantages and appropriateness of semi-structured interviews were summarised by Wellington (1996:21-38) and Arksey and Knight (1999:7-9). Both found that this technique is often the most valuable, offering flexibility within a loosely defined framework. Since the research questions had teacher attitudes as a key feature it was important that the interview technique gave opportunities for clarification or extension of answers – to explore areas of interest that emerge. As a participant researcher, the 'sterility' of a structured interview would be inappropriate with colleagues. At the other extreme an unstructured interview with the Deputy Head could appear too intrusive and may have resulted in poorer responses. Unstructured interviews were also unsuitable in that they are hard to analyse reliably and it can be difficult to ensure anonymity (Arksey and Knight 1999:9). The need to match value-added profiles with teacher responses demanded a more structured approach but not so constrained that depth of responses was compromised.

An issue to be considered was whether or not the interviewee should be told of the results of the quantitative work. To inform an interviewee who, for example, might be less effective, could introduce bias in the data (such as defensiveness) and thereby affect the quality of the data. Conversely, not to tell the interviewee might result in a poorer focus on the key issues and hence less richness in the data.

In order to evaluate the interview schedule, three pilot interviews were conducted. Following these and a series of discussions with a number of experienced colleagues, a 'diplomatic' approach was taken where the interviewee was told the class results - but not immediately. Whilst staff were generally aware of the actual GCSE results of pupils that they had taught, experience suggested that there would be much less awareness about results in value-added terms. As the interview schedule shows (Appendix 1), the initial focus was on each class without the value-added data. Later in the schedule the opportunity to discuss class results was given. This afforded the chance to check initial responses against subsequent ones. For example, a less effective teacher might initially have indicated that relations with a class were good but when presented with the data may have admitted that difficulties were encountered.

The schedule invited the interviewee to relax by asking 'ice breaker' questions about his or her career to date. This also provided information about why the interviewee entered the profession and what attracted them to it. The topic of value-added and value-added policy was introduced, and the teacher was

asked about his or her experiences with two different sets of pupils. The discussion then moved to focus on the Green Paper. This created a deliberate break in discussion about the teaching of particular pupils. The schedule then returned to the identified groups and information about their performance in value-added terms was presented. The schedule then moved to discuss the notion of effective teaching. Hence the interview schedule aimed to draw out a variety of attitudes and behaviours. Possible areas included pupil characteristics that the teacher liked or disliked, expectations, and attitude towards discipline. Attitudes toward (and whether the teacher used) value-added or target-setting data were established.

The pilot interviews all indicated that this approach would yield rich data. Significantly, at the end of each interview all three teachers commented that they were surprised about the amount that they had said. Further, during the interviews the participants were clearly free to 'open up' and demonstrated this by revealing critical incidents in their lives, how they really felt about certain pupils (or colleagues), and in the passion with which they talked about their lives and work. Clearly they also demonstrated trust in me - in every case their senior manager.

McCulloch *et al* (2000:122) confirm that in this sort of interview the emphasis is on validity rather than reliability. As the schedule is not highly structured it will be more difficult to generalise from the interviews and, as McCulloch *et al* found, it would be 'futile to use sophisticated statistical methods to describe the responses' (ibid. p.122). Further, as Hargreaves (1996:13) argued, it would be

inappropriate to condense the voices of all the teachers into a 'singular voice, *the* teacher's voice' (original emphasis). However, it should be possible to make general statements of the type that Bassey (1999:52) calls 'fuzzy generalisations'.

Two other interviews were planned. These were with the former Headteacher and the former Senior Teacher who was responsible for the introduction of YELLIS and ALIS to the school. These interviews were conducted to provide historical evidence covering the period from 1994 to 1997 which will be discussed in chapter 4. There was no formal interview schedule for these interviews since the question was simply 'tell me about the history of value-added at Dalestone School as you remember it'. Since I was also employed at the school at the time it was expected that a complete and accurate account would result.

Interviews were transcribed as necessary; sometimes parts and sometimes a full transcription depending on the perceived quality of the data. The tape record was retained.

Data analysis followed the logical process suggested by Bassey (1999a:69-72). This involved application of Glaser and Strauss' 'grounded theory' (cited by Arksey and Knight 1999 p.162) in that the data items from the first few interviews were interrogated to establish analytical categories or statements. These were refined in the light of subsequent interviews. The emerging concepts and theories were tested against further data. Thus the analysis was

interwoven with the data collection (Arksey and Knight 1999:162) in an iterative process until there was confidence that the analytical statements were trustworthy (Bassey 1999a:71).

In keeping with the principle suggested by Hargreaves (1996:16), one aim in the analysis was be not merely to present teachers' voices but to '*re-present* [Hargreaves' emphasis] them critically and contextually'. He noted that:

> ...there is mounting evidence that what secondary school teachers voice about students, learning, ability, discipline, and subject matter is strongly influenced by the subject communities and academic departments to which these teachers belong (p.16).

Further:

> How a school (or department) is structured, how it is led, and the kind of culture it has developed over time also exerts a contextual influence over how teachers view their students, their colleagues, their work and their own efficacy (p.16).

Therefore it was important in the analysis and subsequent interpretation to focus on what the interviewee is saying, and to review each tape or transcript for a sense of the whole. Categories of analysis were developed which were related to the key research themes, for example, the extent to which teacher attitudes related to differential 'outputs'. It was anticipated that the data would be used to develop constructs of teacher styles, attitudes and behaviours that could be related to differential effectiveness.

2.4.4 Collection of additional school-based evidence – data triangulation

Additional data was necessary to 'map out, or explain more fully, the richness and complexity of human behaviour by studying it from more than one

standpoint' (Cohen and Manion 1994:233). Papers from the Local Authority and school level documents illustrating the development of policy on performance statistics, value-added and target-setting were studied. This included minutes of governors' meetings, reports and memoranda. Documents were also available at departmental level, for example development plans and setting policies. These and evidence from the Headteacher and other senior staff further illustrated the school's direction and expectations and thus added 'completeness' to the data (Jick, 1983 cited by Arksey and Knight 1999:21). In addition to providing factual evidence, these documentary sources provided valuable data about the attitudes and position of key players in the development of value-added policy at the school. These particularly include middle managers (Heads of Department), governors and members of the Senior Management Team (SMT).

An important source of evidence was my observations. This was principally to 'confirm' (Arksey and Knight 1999:21) the interview data or challenge it. This evidence took two forms:

- ◊ Evidence from conversations with colleagues that could be matched against the comments made by colleagues during interviews.
- ◊ Evidence from direct observations of teachers in their classrooms - in particular this took the form of 'through-the-window' observations.

As Deputy Headteacher, I regularly spend up to four (or more) hours per week patrolling the school ('on-call') during lesson time. I am able to observe what is going on in classrooms; I frequently examine pupils' work; hear comments from pupils about staff; read reports to parents; inspect worksheets; read work left on blackboards; read pupil planners (diaries in which, for example, homework is

recorded) and deal with discipline issues. Staff generally expect that 'on-call' teachers will check classrooms, help with discipline issues or simply enter the room to briefly observe lessons in progress.

There were ethical problems associated with this in that my observations could have been interpreted as covert research. Woods (1996:64-65) acknowledged that 'consent' is not a straightforward process and that public and private spheres are not always sharply defined. Burgess (1985:158) commented that both 'open' (where researchers tell the truth and inform individuals about what they doing) and 'closed' (covert) research methods present 'seemingly intractable' problems to which no solutions are instantly available. He suggested that fieldworkers arrive at some form of compromise 'whereby the impossibility of seeking informed consent from everyone, of telling the truth all the time and of protecting everyone's interests is acknowledged' (ibid. p.148). Burgess, for example, reported on conversations that he had with teachers who freely acknowledged difficulties with a group of pupils in the staffroom but who relayed a completely different picture to the Head of the school (ibid. p.152). Burgess was involved in a lie on this occasion, and in his account gave instances where he had been parsimonious with the truth or had asked naïve questions to which he already knew the answers.

I expected to be faced with a similar set of dilemmas. A teacher could make a comment to me that I knew to be an exaggeration or even a lie as a result of my own observations, for example when 'on-call'. For example if a teacher told me that he mainly used group work activities in his lessons yet was mainly observed

using 'chalk-and-talk' then clearly the balance of that data shifted. I thus felt in a strong position to act as a referee in respect of comments that colleagues made and to make the compromises to which Burgess referred.

I felt able to justify this aspect of the research in three ways. Firstly, my observations did not impinge on the subject's right to privacy in the sense that my presence around the school and in classrooms was expected and was part of my professional role. Secondly, this aspect of the research was essentially for the purpose of checking and triangulation. Thirdly, and most significantly, Woods (1996:65-66) discussed such research in terms of 'moral conflict' and the relationship between the researcher and the subjects. He emphasised the need for reflexivity that involves a constant monitoring of the rightness of what one is doing. In this research I am confident of good relationships and a clear conscience.

Some of the opportunities and difficulties experienced by Pollard (1985) as a full-time teacher doing an ethnographic study in his own school were anticipated in this research. He was conscious of the fact that he had to maintain both the role of a participant and that of observer in an 'unending dialectic' (ibid. p.219). Pollard emphasised the need to achieve a degree of detachment in order to record and analyse data effectively – that is, to maintain the 'counter-balance of the researcher perspective' (p.220). He mentions several factors that were important in his research including a good relationship with the Headteacher and with other staff; and being a good teacher to develop 'subsidiary awareness' (p.221) and acquire 'tacit knowledge' (p.232). Being primarily a full-

time teacher was highlighted by Pollard as a significant feature. He was in a different position from most participant observers whose participation forms part of the research strategy alone. Pollard's participation was 'real' but this inevitably posed problems. He discussed the issue of time and the need to meet demands of the full-time job. More significant were ethical problems associated with collecting data from the staff. Pollard's description of 'semi-covert' methodology (p.225) seems to be useful in the context of this study. Colleagues were initially aware of the research focus and *elements* of the data collection process. This issue was resolved later in the interview process and in the reporting back to colleagues – that is to say, the covertness was discarded later. As Pollard said:

> This was a pragmatic solution to a difficult dilemma which involved not only the question of ethics but also that of the quality or 'naturalness' of data collected (p.225).

2.4.5 Positionality - The Researcher as Deputy Headteacher and a Parent

Thompson (1999:1) stated that the duality of role of researching professionals presents particular challenges when 'own patch' research is carried out. He considered that a critical step in meeting these challenges was to develop a high level of self-awareness in research 'through an appreciation of broad *positionality* as defined by the social, cultural, political, professional and other factors that shape their personal ideograph' (ibid. p.1, Thompson's emphasis). In my case there was a triplicity of roles - researcher, deputy headteacher and parent - and the notion of positionality demanded that I considered and discussed personal influences on the research process.

Burgess (1984:210) recognised that in research 'the biography of the individual researcher has a part to play'. Thus I have been involved in and enjoyed value-added work at Dalestone since the school joined ALIS and YELLIS. I enjoy working with numbers and looking for patterns in data. As a Science Teacher I have yet another role and have had a personal interest in the performance of my classes and have frequently reflected upon the part that I have played in the outcomes.

Considering the scientist in me, it thus seems natural that there should be a quantitative element in this research. The qualitative elements originate from my assumed role as a senior manager as a 'people person' which itself is undoubtedly influenced by my Christian faith.

I work full time as Deputy Headteacher at this school. My relationships with staff are excellent (independently judged by an assessor from Leeds University for the National Professional Qualification for Headteachers) and I have their full co-operation. In consideration of my position, interviewing was reflexive rather than standardised (Hammersley and Atkinson 1993:112). Research in the school where I work raised particular ethical issues, for example the interviewing of a subordinate colleague whose 'effectiveness' as revealed by value-added data may be 'negative'.

It was clear that the research also raised ethical issues for the interviewees too. The quality of relationships between colleagues and myself were so good that

they sometimes entered 'territory' that might be considered to be 'unprofessional' when talking to a Deputy Head. In several cases the conversation stopped and gestures were made which would not be recorded. Such actions actually add to the richness of the data presented in this study and are of crucial importance in symbolic interactionist research (Woods 1996:41).

Colleagues were guaranteed confidentiality and their contributions have been made anonymous. The potential advantages and problems of practitioner research were summarised by Wellington (1996:15). For me the main advantages were access and good relationships with colleagues. The chief obstacle for a Deputy Head is always time!

Being the parent of two children at the school potentially added a further complication to this research. However, neither child was in any of the classes to be discussed with colleagues as they were too young. It was also fortunate that both have excellent relationships with their teachers and as a parent I have been satisfied with the quality of teaching that they are receiving.

2.4.6 Final Analysis

Following analysis of the interview data the next stage was to match the outcomes from the interviews with the profiles of value-added scores achieved by each teacher. Since the interview schedule was closely matched to the research questions it was anticipated that the analytical categories that emerge would align with the value-added data. Thus, for example, teacher statements in

categories about 'expectations' was matched with the actual value-added scores achieved by the sets that those teachers taught.

The outcomes of the 'matching' process were refined in the light of the additional evidence provided by documents, my observations and reflexion. In this way generalisations that address the research questions could be formulated.

Davies and West-Burnham (1997:223) commented that

> One of the greatest fallacies of late twentieth-century educational policy-making at macro and micro levels, is that there is a necessary and contingent relationship between teaching and learning. Teaching can take place without any learning resulting from it and a great deal of learning is achieved without the benefit of teaching.

Although there is undoubtedly truth in Davies and West-Burnham's second sentence, the first is an overstatement. Evidence from the school effectiveness and school improvement movements emphasise that 'schools do make a difference' (Hopkins, Ainscow et al. 1994:43) and go on to demonstrate that the use of appropriate teaching strategies can dramatically increase student achievement (ibid. pp.51-67). In their pursuit of performance-related pay for teachers, the government clearly also accept a link between teaching and pupil outcomes. Thus it does not seem unreasonable to search for generalisations that connect teaching and value-added outcomes.

When I started this study, the scientist in me was constantly calling for a methodology in which the research outcomes could be subjected to statistical tests and subsequently clear-cut conclusions stated. As work has progressed I

have reflected on a comment by Fitz-Gibbon (1996:167) that effectively warns about the arbitrary nature of 'statistical significance' and stresses that the real need is to look for what is 'educationally important'.

2.5 Concluding comments

One of the most significant outcomes from this methodology has been the sheer volume of data that it has generated. Probably the most important reason for the wealth of numerical, oral and documentary evidence has been the fact that I have conducted this research not only as a participant researcher with 'access to answers that would have been denied a complete outsider' (Arksey and Knight 1999:11) but that I am in a senior position with years of experience, unrestricted access to information, tacit knowledge about all aspects of the school, and the confidence of colleagues. The richness of interview data, for example, demonstrates that any initial reservations that I had about whether teachers would 'open up' were unfounded. As a consequence, the chapters that follow offer 'layers of reality' (Woods 1996:38) which are much deeper than the 'public face' of Dalestone School.

Chapter 3 The raw and the cooked: towards a 'fairer' way of using performance data?

> *How can 'fairer' comparisons be made between the results of individual secondary schools? This is a question which has preoccupied researchers of school effectiveness over much of the past decade. It is one, however, which has assumed still greater importance in the light of the recent requirement of the 1980 Education Act, in a climate of demands for greater 'accountability' in the public sector, that schools should publish their examination results.... Our answer [to the question about whether 'fairer' comparisons can be made] is unequivocal: they can be. (Gray, Jesson et al. 1986: 91, 116)*

3.1 Introduction

The purpose of this chapter is to chart a critical path through published literature in order to trace the development of value-added policy and practice. The difficulty of this task is not underestimated. Finkelstein noted that:

> In order to explore policy-practice connections, historians would have to integrate studies of teacher practice, student culture, curriculum, and assessment processes into their explorations of contexts, regulatory authority, market realities, and national ideologies as they changed over time (1997:309)

Accordingly, this chapter will examine the political and policy developments that have given rise to the need for a 'fairer' way. Consideration will be given to the relationship between policy and practice and the context in which value-added had developed. This account will examine the utility of value-added research particularly with regard to the accurate identification of effective schools and teachers. Accountability has increasingly been the watchword of education. Schools are required to account for their performance and this inevitably requires objective measures so that one institution can be compared with another. This, together with the marketization of education, for example in

increasing parental choice, had led to a focus on examination results as the measure of a school's 'output'.

A recurring theme is the debate about how the examination data should be used in the determination of performance. Some prefer it 'raw' whilst others favour statistical treatment - although the use of some techniques has led to comments about results being 'half-baked' (Goldstein 1997a) or masking the truth, that is, 'cooked'. Value-added, with its roots in school effectiveness research, was originally conceived as a tool in furthering our understanding of the complexities of educational structures and how they influence pupil outcomes. However, it has come to be recognised by many as a fairer way of identifying 'effective' and 'failing' schools.

Whilst considerable attention remains on individual schools it is to the performance of departments and individual teachers that the focus has now moved. The revelation that the teacher 'level' explains up to four times more in terms of pupil results than the school 'level' (Coe and Fitz-Gibbon 1998:429 ; Reynolds 1999b) has brought the classroom under close scrutiny. Thus in the existing climate of accountability the 'holy grail' has become the identification of the 'effective teacher'. This pursuit has resulted in a confused and labyrinthine situation permeated by politics, statistics, pedagogy and culture. The research community has not found a way out of this since research into effective teaching has been approached in a number of different ways and there are many different perspectives on the theme of effective teaching (Harris 1998:170).

Yet our knowledge of teacher effectiveness is *not,* as Reynolds asserted (1998b:26), 'virtually total'. In her overview of the research findings concerning effective teaching, Harris (1998:169) found that the relevant research literature is 'both vast and complex'. Hay McBer found that in over 200 separate studies there was little consensus about how to define the qualities of an effective teacher (BERA 2001:7). Thus the problem of identifying the 'effective teacher' lies in that very complexity which Reynolds has been accused of trivialising (Edwards 1998:29).

Complexity is a key issue in this study. Nuttall *et al* (1989:776) stated that 'to attempt to summarise school differences even after adjusting for intake, sex and ethnic background of the students and fixed characteristics of schools, in a single quantity is misleading'. They were acknowledging a phenomenon known as 'differential school effectiveness' whereby a school might not be equally effective for all groups of pupils, for example boys, girls, ethnic minorities or even one subject compared to another. 'Setting' is intimately associated with differential school effectiveness and questions are frequently raised about the processes used to generate sets. The representation of different groups (for example ethnic minorities) in top versus bottom sets and the efficacy of setting in raising achievement are important. The work of Jo Boaler is significant here - her conclusion from a recent study was that setting 'could be the single most important cause of the low levels of achievement in mathematics in the UK' (Boaler, Wiliam et al. 2000:646).

The chapter concludes with a summary of the current position of value-added systems. It was largely the reaction of schools to the imposed publication of raw examination results in the 1980s that led to the growth of such systems. Saunders (1999:234) commented that 'you could say that value-added was an idea waiting for its time'. Value-added systems are seen as being fair in that they take into account the prior attainments of pupils and show the progress that they have made in the school.

3.2 Policy and practice

There is a dynamic, reciprocal relationship between teachers and educational reforms. Reforms are seen to affect teachers but teachers also affect the reforms. Ball, Helsby and McCulloch are amongst educational researchers who have recognised the importance of teachers' interpretation or mediation of government policy.

Ball (1994:180) reasoned that policies differ in their form and forcefulness. He cited the National Curriculum as having legal force, Local Management of Schools (LMS) as creating a tightly defined administrative framework, and other policies as being 'contradictory and vague like the SEN responsibilities of LEAs'. As a consequence, it follows that the interpretations of policies will vary. Thus, in a study of the National Curriculum, Bowe *et al* found that:

> [It] is not so much being 'implemented' in schools as being 're-created, not so much 'reproduced' as being 'produced'. While schools are changing as a result, so too is the National Curriculum...[It indicates that the power of the State] is strongly circumscribed by the contextual features of the institutions, over which the State may find that control is problematic and contradictory in terms of other political projects. (Bowe, Ball et al. 1992:120)

They argued that the 'implementation' of policy is 'subject to interpretational slippage and contestation' (Bowe, Ball et al. 1992:83). Hatcher and Troyna (1994:163) agreed but emphasised that although the State cannot impose an interpretation at the level of discourse it 'certainly *can* impose one at the level of practice' (original emphasis). The National Curriculum framework was 'loathed and resented by many who saw it as a government attack on teachers' autonomy, integrity and professionalism' (*Times Educational Supplement* 1998a) but it *is* in place. Similarly, schools now operate within the framework of LMS.

It is crucial that the complexity of the relationship between policy and practice is appreciated. Four important aspects of the complex and contested reality of the relationship between teachers and the State are examined by McCulloch *et al* (2000:62). They stated:

- ◊ that "professionalism" is not a clear-cut concept
- ◊ that the idea of occupational autonomy is problematic
- ◊ that teachers are influential in educational policy - particularly in its interpretation
- ◊ and that a reductionist view of the curriculum ignores aspects that are relevant to any consideration of teacher professionalism (2000:64).

In another publication, McCulloch (2000) demonstrated that the growth of competitive pressures (going back to the 1950s) surrounding examinations rather than the National Curriculum *per se* eroded teacher freedom in the curriculum. At the same time, Helsby (2000) concluded that the initial implementation of the National Curriculum had a broadly negative effect on teacher professionalism. Significantly, she distinguished between control of

curriculum content and control of pedagogy and suggested that the latter lay at the heart of teacher professionalism (ibid. p.96). Helsby (ibid. p.106) subsequently pointed to a number of factors leading to a diminishing of teachers' professional confidence such as top-down accountability, the intensification of working life, resource and time constraints. Fundamental contradictions in government policies are seen to underpin these factors. Despite pledges to improve the quality of education, funding is reduced or, more subtly, flowing from those children with the greatest need to those with the least need under the market system and the 'concomitant processes of decomprehensivization' (Gewirtz, Ball et al. 1995:189).

Black and Wiliam (1998) viewed this mismatch between policy and practice somewhat differently. In an examination of one aspect of teaching - formative assessment - they state that the policy perspective is wrong and ineffective in the first place. In particular they believe (p.1) that government policy has seen the classroom as a *black box* - that is, it expects that standards will be raised but ignores what goes on in classrooms; 'it is up to teachers - they have to make the inside work better'.

The New Labour government envisaged a 'new professionalism' with better rewards and support in return for higher standards - effectively defined in terms of examination results (DfEE 1998). Chitty and Lawn (1995:141-2) suggested that the influence of teachers should not be 'written off' and McCulloch *et al* questioned whether the improvements that the government required would happen without the fullest co-operation of teachers:

> Neither raising standards by regulation nor professionalizing by prescription will work. Teachers have power in the sense that they have to want improvement for improvement to happen (2000:118).

However, empirical research by Halpin *et al* (1999:12) suggested that whilst many teachers and heads were embracing change, it was not always in the ways and of the kind of which government might approve. Rather 'it is through processes involving selective acceptance, partial rejection and varying degrees of appropriation and synthesis' (ibid).

2.1.1. Value-added policy and practice

Value-added policy and practice has developed in the above context through a number of significant changes in direction. As a critical issue, value-added originated in school effectiveness studies in the USA as early as 1970. It was not until the late 1980s and in particular the Education Reform Act (1988) that the centrally-imposed funding-and-accountability mechanisms introduced a rationale for schools to be concerned with the idea (Saunders 1999:238).

The pioneering work of Fitz-Gibbon's ALIS project at Newcastle together with the work of others such as Goldstein, Gray and Jesson was timely. Saunders (1999:246) believed that the strength of the pioneers' conceptual and methodological frameworks enabled them to challenge the government's proposals to publish the results of national testing promptly and capably. Goldstein and Cuttance (1988:201) for example, urged the government to reconsider introducing a system which 'is likely to result in the unjustified victimisation of schools in socially disadvantaged communities'. Similarly, Nuttall

et al (1989:770-776) warned about summarising school differences in a single statistic particularly in view of differential school effectiveness and questions over the stability of school effects over time. Subsequently, a briefing paper for the National Commission on Education by McPherson (1997) acknowledged that 'raw' outcome measures demonstrate actual attainment but clearly explained that 'an assessment of the contribution a school makes to its pupils' progress' was needed in a 'good indicator system'.

The research community's arguments were effectively ignored as the government promoted performance tables using 'raw' data. Schools, however, were convinced by the value-added argument and the two Headteacher Associations (the National Association of Headteachers, NAHT, and the Secondary Heads Association, SHA) promoted the CEM Centre's Year 11 Information System (YELLIS) (Curriculum Evaluation and Management Centre 1994). YELLIS grew from 49 participating schools in 1993 to 242 in 1994.

Hence by 1994, government, and notably the then Secretary of State for Education, Gillian Shephard, was persuaded to consider the value-added concept and a prolonged period of research, pilot investigations and consultation followed. More details of this period are given in the next section.

Yet a change of government and seven years further on have not significantly moved policy as far as performance tables are concerned. Many researchers such as Fitz-Gibbon (1997), Goldstein and Thomas (1995:37) and Saunders (1997) have questioned the use of value-added analyses for the purpose of

public accountability mainly on the grounds that such measures are not sufficiently simple and straightforward. Thus although many schools use value-added analyses, government policy is vague and weak on this but resolute in the use of the performance tables that use 'raw' data.

The 1997 Education Act provided legislation requiring targets to be set and published by the governing bodies of maintained schools. This marked a significant policy shift not away from value-added but rather by opening up a new strategy aimed at raising standards. This involved the use of the same sort of baseline data to that used in value-added analyses, but schools were now required to compare their performance with other schools using benchmark information provided annually in PANDA and Autumn Package reports. Schools can use these reports to construct simple 'value-added' graphs for internal consumption. Thus, as with National Curriculum policy, so target-setting policy is now in place.

The most recent policy shift has been described as 'turning the tables on teachers' (Hardy 1998:7). Hardy laments the delay in including a value-added element in league tables but notes that the green paper *teachers: meeting the challenge of change* includes a Threshold Assessment which includes a measure of pupil progress. The Threshold Assessment pack (DfEE 2000) suggests that value-added data can be used in the application for a £2,000 pay rise. Threshold Assessment and Performance Management policies are thus in place. One of the issues that this study will explore is how they will impact on practice.

3.3 The context for the development of a 'fairer' way

The origins of value-added in an educational context can be traced to school effectiveness research in the United States of America in the 1960s and 1970s. At this time much research in the US examined the extent to which race, IQ, socio-economic status (SES), and individual schools determined educational achievement (Sammons 1999:25). Sammons reported that this work had an emphasis on equity and a focus on raising the achievement of poor and ethnic minority students. Although two national indicators were developed - the 'Equality of Educational Opportunity Survey (EEOS)' and the 'National Assessment of Educational Progress' - no full-blown national indicator system has resulted largely because the authority and responsibility for education lies with the states (Fitz-Gibbon and Kochan 2000:274). Significantly, the EEOS (otherwise known after its principal author as the 'Coleman Report') concluded that

> ...socio-economic factors bear a strong relation to academic achievement. When these factors are statistically controlled, however, it appears that differences between schools account for only a small fraction of differences in pupil achievement. (Coleman, Campbell et al. 1966:21-22)

In another American study, Jencks *et al* (1973) reached similar conclusions from a large, though not longitudinal, database. Bernstein (1970:344-5) however, argued that the context in which children were learning was primarily a middle class one. He maintained that education could not 'compensate' for society and so there was a need to 'consider... the conditions and contexts of the educational environment'.

By the late 1970s researchers were moving to conclusions that were opposite to those of Coleman and Jencks. Rutter *et al* (1979) showed that some internal school factors were associated with effective schools that stimulated further studies which showed that 'schools do make a difference' (Hopkins, Ainscow et al. 1994:43). The emergence of this conclusion thus stimulated the idea that some schools make more difference than others and that they should be accountable for this difference.

In his 1976 speech at Ruskin College, Oxford, James Callaghan (then Labour Prime Minister) called for a rise in standards and greater accountability within schools and stated 'if the public is not convinced then the profession will be laying up trouble for themselves in the future' (1976:14). Since that time 'accountability' has increasingly become the watchword of government education policy through, for example, the introduction of performance tables, the establishment of OFSTED, and most recently the introduction of Performance Management. In his Ruskin speech twenty years after Callaghan, the present Prime Minister, Tony Blair (then Leader of the Opposition) promised 'zero tolerance of school underperformance' (1996).

What is the background to this political preoccupation with standards and the wish of politicians to get better value for public expenditure? Saunders believed that it is about economic advantage:

> For some time this agenda has been explicitly attached to the issue of global competitiveness and the belief that education and training are important levers for economic competitive advantage. (1999:234)

However in contrast, Barber (1996:19) proposed that politicians' greater interest in education is a result of their *loss* of control of the 'macro-economy levers that they had previously used to achieve their goals'. Instead, he suggested, they now use education policy levers. Wider attempts to reduce public expenditure, reform public services and develop individual consumer rights also form a partial explanation. Deem (1994:24) suggested that changes partly emanate from the desire of 'New Right' politicians to expose publicly funded social institutions to the 'market place' and this is evident in the devolution of budgetary and management responsibilities. Coupled with this would be the suggestion that recent governments have been determined to reduce the influence of professionals at all levels throughout the education service. In 1988 Sallis (p.135) for example, observed that Sir Keith Joseph (Education Secretary at the time of the 1986 Education Act) saw governing bodies as a sort of consumer council holding professionals to account.

It took time to convince politicians about the emerging value-added models. Saunders (1999:235) reported that the Secretary of State for Education (1992-4), John Patten, was 'scathing' about the development of ways of 'cooking schools' results'. It is true that he preferred raw results:

> It is too easy to point a finger and say glibly of examination statistics that they are raw results. If they are raw results, young people carry those results with pride in their pockets as their educational currency throughout their adult working lives. They are the things which employers wish to see. (Patten 1993).

However, later in the same Commons Speech he acknowledged the need for research into value-added:

> I do not believe that any of us should delude ourselves. We would all like to see more value-added measures in the performance tables, but it will not be an easy task to get agreement on what those measures should be. A lot of work must be done. [ibid]

In 1992 the Conservative Party secured a fourth term in office, albeit with a reduced majority. This was however, sufficient to continue their programme of education reforms. 1992 to 1997 was a formative period for value-added. A well-prepared research community had met the publication of league tables that used 'raw' data. Schools were embracing the idea of value-added and there was rapid growth in systems such as YELLIS.

In 1992 the Conservative government published the White Paper *Choice and Diversity* which particularly aimed to increase parental choice and school autonomy, and to secure greater accountability. Significantly, there is no mention of value-added in this document. However, the White Paper established the School Curriculum and Assessment Authority (SCAA) and stated that one objective was to 'put governing bodies and headteachers under greater pressure of public accountability' (DFE 1992:18). Of interest is the final chapter where in the context of raising standards the word 'benchmark' is used twice (ibid. p.64). It is suggested that this is perhaps a pointer to policy away from value-added and towards what is now in place in the form of the PANDA report and Autumn Package.

Nevertheless, the idea that 'value-added' might give a clearer impression of a school's performance was eventually accepted by the government. A SCAA working party reported on the advantages of 'value-added' in 1994 but Gray (1994) was disappointed that the multilevel model was initially rejected in favour

of discredited, erroneous or 'back-of-the-envelope' models. Amongst their recommendations SCAA (1994) said that more research was needed. Fitz-Gibbon's team was commissioned by the Government to conduct research and advise the Secretary of State on the development of a national system of value-added reporting for schools. The Department for Education issued a briefing paper in 1995 to demonstrate its commitment to value-added stating that:

> This is a better reflection of schools' achievements than raw performance tables since the effects of socio-economic factors will be largely cancelled out (DFE 1995:2).

With their election in 1997, the Labour Party published a White Paper, *Excellence in Schools*, which set out their Labour education agenda for the next five years. Gibson and Asthana (1998a) and Goldstein agree that the new government had 'hijacked' (Goldstein 1997b:395 ; Goldstein 1998:521) the findings of school effectiveness research for political purposes. The White Paper stated that national measures of pupil achievement showed that:

> ...in practice, schools with similar intakes of pupils achieve widely differing results. The differences are a measure of a school's effectiveness in teaching and motivating its pupils. (DfEE 1997:25)

The conclusion was clearly the sort of abuse of the findings of school effectiveness research that Reynolds has condemned (Reynolds 1998a). It was also a political statement which illustrated Gibson and Asthana's assertion that the White Paper reflected the survival of 'a perspective more in keeping with New Right thinking' than what might be expected from a Labour party which traditionally has a concern for equality of opportunity (Gibson and Asthana 1998a:195-6). They argued that the Conservative government believed that school improvement would take place through the pressures of competition that

were generated by the 'quasi-market in education' whereas under Labour the pressure would be applied through performance targets (ibid. p.196). Crucially, both regimes placed the onus for improvement directly upon schools:

> All the evidence indicates that standards rise fastest where schools themselves take responsibility for their own improvement (DfEE 1997:24).
>
> The main responsibility for improving schools lies with schools themselves (ibid. p.12)

In an examination of their education policy, Power and Whitty (1999:537) found that New Labour has sought to control education more directly. They concluded (p.539) that the government had refined rather than rejected the principles of quasi-markets, thereby continuing a politically-driven emphasis on school-level accountability. They pointed to more sophisticated league tables and the introduction of 'target-setting' as examples of this.

The Value Added National Project Final Report (Fitz-Gibbon 1997) was followed by a SCAA consultative paper (1997) which made only two main recommendations to the government. The first was that value-added feedback should be provided to schools for internal school management, and the second was that further feasibility studies should be carried out. The successor to SCAA, the Qualifications and Curriculum Authority (QCA), produced a further consultation (1998) and the New Labour government intended to include a 'progress index' in the Performance Tables that autumn. This was hastily dropped after complaints that this index was not an accurate reflection of the effectiveness of high achieving schools. In a compromise only those schools which had made progress against earlier results appeared with a tick against their name in the tables. Academics had serious concerns about the methods

used in the pilot. Fitz-Gibbon reiterated her concerns over the use of a single figure to describe a school, whilst Goldstein criticised the pilot's 'crude and simple' mathematics (*Times Educational Supplement* 1998b). In 1999 the progress index was dropped but a value-added pilot scheme involving about 200 secondary schools is running and it is likely to be 2003 before any national value-added league tables are published.

It is clear that there has been a legacy of delay on the part of government in respect of the implementation of any national system of value-added. Some of this hold-up is justified and was in fact recommended by Fitz-Gibbon (1997:95-6) particularly where value-added data would be used for public accountability. However, the recommendation that publication should be based on at least three years' data to take into account year to year instability is expected to be sidelined in favour of tracking pupils from Key Stage to Key Stage. The government also recommend that issues raised in the *Value Added National Project Report* such as pupil mobility, need to be dealt with before reporting value-added data (SCAA 1997:13). Yet, as Goldstein (1997a) correctly pointed out, if the government 'accepts such reservations about value-added tables then logically it ought to do so for the current tables'.

Further, it seems that policy-makers are failing to take into account research which shows that schools do change in their effectiveness but that this is not a fast process. For example, in their conclusion Gray *et al* (1996:50) said:

> Our evidence suggests that it would take several years for a relatively ineffective school to get into the pack of schools deemed to be initially of average effectiveness and as long again for them to pull ahead.

Only the recommendation in the *Value Added National Project Report* (ibid. p.97) to introduce a unique pupil identifier received an immediate and positive response. Rather than develop Fitz-Gibbon's ideas for the internal use of value-added, the DfEE has concentrated on its Autumn Package and PANDA reports. The former facility allows schools to produce 'value-added graph plots' whereby a set of pupil results can be compared with national data but not, for example, the ability to compare one subject directly with another which a system such as YELLIS permits. In the PANDA, a school graded as, for example, 'E' (well below) in comparison to the average for similar schools for 5 or more grades A*-C at GCSE will always remain an 'E' if all of the schools make the same amount of improvement. The direct value of this sort of analysis is therefore questionable. However, since this set of data is published in, for example, OFSTED Inspection reports it could be argued that its utility lies in the pressures that it creates within schools for improvement.

Yet a system of national value-added league tables is likely to attract as much opposition as those using raw data. Research by David Jesson showing that grammar schools 'add less value' provoked an angry response from the National Grammar Schools Association (Dean 2000; Hale 2000). Similarly, claims that 'secondary moderns do best of all' were said to have 'muddied the waters again' (Barnard 2000b). Fitz-Gibbon considered that league tables (1996:73) take no account of the kind of students in the school, that results do not tell the whole story of a school's quality for example in sports, music or quality of relationships. Referring to raw data, Fitz-Gibbon suggested that performance tables have led to an increase in exclusions (1998:24). It is suggested that schools would be

just as keen to remove badly behaved pupils where their standing in a league table based on value-added measures was threatened. Further, she rightly stated that whole school aggregates are of almost no use for improvement purposes - fine-grained data would be needed at class and teacher level where much of the interesting variation occurs (for example see Gibson and Asthana 1998a). However, such data is far too sensitive to publish since it will relate to the work of individual teachers.

3.4 Statistics - the uncertainty principle

Although not referred to as 'value-added', the use of residuals as a tool in educational research had been developed by 1970. Rosenshine (1971) reported on over 50 studies in which some measure of teacher behaviour was related to one or more measures of student achievement. In the majority of these studies the researchers calculated 'residual gain scores: that is, the difference between the actual post-test score and the score that would have been expected on the basis of the initial score' (p.24). However, the use of a variety of techniques and the absence of a common baseline measurement significantly reduces the value of this work.

The research community quickly perceived unfairness in the ways in which examination data was being used to judge schools. Gray and Hannon (1986:23), for example, examined the extent to which the strategies employed by HMI were equally fair to all schools. They found HMI judgements of schools' examination results to be unsystematic but a realistic response to the variety of

data and opportunities for analysis available. Gray, Jesson and Jones (1986) considered what a 'fair' comparison of the performance and effectiveness of schools would look like and then examined prevailing approaches. They found that although comparisons between secondary schools were essentially confined to examination results, there was a variety of approaches in use and misinterpretations of the statistical evidence.

The lack of 'tools' by which to fairly judge the performance of schools was a key issue and researchers set about this task. The work of Gray *et al* was significant in that they found a 'mismatch between the systematic efforts of individual LEAs to collect examination results on one hand, and the absence of comparable efforts to collect data on schools' contexts on the other' (ibid. p.99). Having also recognised their 'emerging understanding of statistical issues' (ibid. p.99) they changed their research strategy and this led to the development of a number of statistical models in an attempt to capture schools' effectiveness.

The paper by Gray *et al* was most significant in that the authors highlighted a number of issues that have been important in the development of value-added, for example the need for pupil-level data and reliable prior attainment measures. It should be noted that at this stage such research work had comprised a series of innovative projects with no particular policy focus (Saunders 1999:245).

In 1983 the A-Level Information System (ALIS) project was established at the Curriculum, Evaluation and Management Centre (CEM) at Newcastle (latterly Durham) University under the direction of Carol Fitz-Gibbon. It was not until

1992 (with the publication of performance tables) that the work took on a new urgency. One of the first uses of ALIS was to try to explain the perceived variation in examination severity across Exam Boards and subjects. Tymms and Fitz-Gibbon (1991:31) concluded that there was 'an impressive degree of consistency amongst the Boards'. Their suggestion that there might be differences (in terms of difficulty) across subjects was judged to be important by schools. Subsequent research confirmed their hypothesis.

Harvey Goldstein has been one of the pioneers of value-added research but has also been one of the most vocal about the pitfalls that can be encountered when it comes to interpreting statistics (see for example Goldstein and Thomas 1995; Goldstein 1997b; Goldstein 1997a; Goldstein 1998). In particular he pointed to 'serious and *inherent* limitations to the usefulness of performance indicators for providing reliable judgements about institutions' (1997b:370, emphasis in original). He noted that there was a tendency to use performance indicators to rank schools rather than to explain their differences. Even when differences were found to exist:

> ...we cannot, with any useful precision, decide how well a *particular* school or department is performing: this 'uncertainty principle' operates to provide a fundamental barrier to such knowledge (Goldstein 1997b:371, emphasis in original).

The Department for Education had already partially acknowledged this in the 1995 briefing paper. 'Inherent difficulties' were stated as a reason why it was unlikely to be possible to measure value-added with sufficient accuracy to put institutions in a simple rank order. However, the DFE did hope to establish whether an institution's value-added was significantly above or below the national average.

Problems of significance were also encountered by Gray *et al* (1995:111-113) when studying changes in value-added estimates over time in order to judge 'improvement'. Also more recently in a study of one LEA, Yang *et al* (1999:480) said that the use of value-added estimates in any form of league table was 'inappropriate and would destroy their credibility and usefulness'. It should be emphasised that the government does not publish 'league tables', rather they produce alphabetical lists of data for individual schools and LEAs. However, the *intention* is that users will make comparisons between schools; and the media and LEAs usually rank the results in order to identify the 'best' and 'worst' schools.

Much of the research into school effectiveness measures the effectiveness once only. Longitudinal studies are relatively rare (Scheerens 1992:9; Gray, Jesson et al. 1995; Sammons, Thomas et al. 1997:48) despite warnings about the unreliability of measures of outcome in, for example, a single year (Nuttall, Goldstein et al. 1989:775). Gray *et al* (1995:98) suggested that this was primarily because researchers have tended to seek to replicate their findings, that is, to look for stability in school effects. They suggested that the converse, instability, is essential for the study of change.

The available data from the researchers cited in the last paragraph suggests that there is a good deal of stability in schools' effectiveness from year to year. However, once again the aggregation of data is inevitably a feature in these findings and 'stability' conceals 'instability' or differential effects within different

subjects or groups within the school (Mortimore 1998:94) - referred to as the 'swings and roundabouts effect'. This has been clearly demonstrated for single subjects by Fitz-Gibbon, Tymms and Hazelwood (1989) over a five-year period and the importance of the class effect has also been demonstrated in Australia by Hill and Rowe (1996). Using a multilevel technique they demonstrated that 'class effects are very large and that the unique effect of schools over and above that due to within-school class differences are relatively small' - that is, schools make a difference but they do so mainly at the level of the class (ibid. p. 26).

It follows that 'instability' effects must be even greater the more disaggregated the data becomes. Tymms (1993:292) demonstrated a general principle that

> ...the smaller the unit of analysis and the closer one gets to the pupils' experience of education the greater the proportion of variance explicable by that unit.

Thus at the level of the individual teacher there is the potential for great variation. This is commonly seen in OFSTED Inspection Reports where, as a result of lesson observations (admittedly limited in number and under abnormal conditions) the teaching in a school overall might be described as 'good'. This can mask a range from 'excellent' to 'poor' - and clearly school managers need to know where in the school each standard of teaching is occurring. It is important also to appreciate that a single teacher who was observed teaching on two occasions, one graded as excellent and the other poor, will, overall, be classified as a 'good' teacher. As has already been stated, the teacher level is the current focus of interest but the fact that data (for example examination results) for an individual or even a small group of teachers are limited means

that statements about the statistical significance of their performance are unlikely to be possible.

Fitz-Gibbon offered a pragmatic approach. She (1996:167) differentiated between 'educational importance' and 'statistical significance'. The latter might use, for example 95% confidence limits, whereas the former is judged by watching the indicators in context from year to year:

> The best way to use indicators is thus light monitoring by those closest to the data (ibid. p.167).

Statistical significance and educational importance might be equated with Hargreaves' (1994:59) notions of 'scientific certainty' and 'situated certainty', the latter being firmly within the context of the school. In an earlier publication Fitz-Gibbon stated that:

> ...management teams in schools must receive the information and respond in ways which are "versatile and flexible". The data must be treated as just one more source of inspiration for "options", rather than as a firm finding, let alone "eternal truth" (Fitz-Gibbon, Tymms et al. 1989:146).

Thus value-added data are best used as part of a quality control system and despite his well articulated reservations even Goldstein (1997b:372) admits that:

> The use of adjusted school or classroom estimates to detect very discrepant units does have certain uses. If handled with care, such data may also be useful as a component of schools' own self evaluation.

A further problem with performance data is that by the time it has been analysed it refers to a cohort of students that entered the school several years previously and may have left before the analysis is complete (Goldstein and Thomas 1995:37; Sammons, Thomas et al. 1997:30). To this must be added the probability that staff and syllabus changes are likely to occur over time.

Together with changes in the nature of pupil intake and numerous other reasons (see Fitz-Gibbon, Tymms et al. 1989:144) there might seem to be little point in doing an analysis in the first place. Goldstein and Thomas (1995:37) and Goldstein (1997b:371-2) however, recommended that value-added estimates are best used to identify 'outliers' for followup research or as a form of screening instrument.

Murphy (1996:32-3) was concerned about the 'outrageous conclusions' that he saw being drawn from the results of national assessments. In a simple but logical way he demonstrated the limitations of educational assessment procedures and particularly found that the results were 'at their weakest when it comes to comparing performance across different aspects of the curriculum or across different years'. He went a step further than Fitz-Gibbon by suggesting that the most appropriate use of assessment data is when it is used 'by and for *individual* pupils' (emphasis in original).

Hence it is the inappropriate use of performance data that has angered many of the prominent value-added research pioneers. In a critical paper, Goldstein (2001:442) has concluded that government has largely ignored researchers' findings about the limitations of performance data in terms of practice, 'even while accepting the limitations in theory'.

A final comment about uncertainty relates to the notion of cause and effect in school effectiveness research. Many researchers have produced lists of features of effective schools or teachers (for summaries see Hopkins, Ainscow

et al. 1994; or Cullingford 1995). In many cases these features are ones which correlate well with outcome data such as examination results or value-added measures. Whilst such correlates can be of use in school improvement it must be emphasised that a correlation does not imply a causation (Scheerens 1992:72; Fitz-Gibbon 1996:104; McPherson 1997:185). Fitz-Gibbon believed that this led to many overstated research 'findings' and cited the common example of 'high expectations' as one correlate of effective schools which is 'often used to simply blame teachers without evidence' (Fitz-Gibbon 1996:106). A BERA methodological seminar found the Hay McBer research design guilty of confusing correlation and causality (BERA 2001:7). Thus in the practice of using value-added data great caution needs to be exercised in drawing conclusions, for example about the effectiveness of teachers.

Similary, whilst there is a strong body of evidence that qualities such as 'high expectations' appear to be important in good teaching, there is other evidence which casts some doubt on such conclusions. Thus Raudenbush (1984:93) concluded that after 14 years' research, teacher expectancy theory was still 'under a cloud of controversy'.

3.5 Reaching the classroom level

> *...schools cannot be represented by...single essentialising tags. (Ball 1997a:347).*

As early as 1986 Gray *et al* (1986: 92) appreciated that 'the circumstances under which "fair" comparisons of performance between schools may be achieved are highly constrained'. They noticed, for example, differences in

schools' effectiveness between more and less able pupils. By the late 1980s it had been established that some schools were more effective than others but Nuttall *et al* raised the question as to whether schools were equally effective for all groups (1989:769). It is clear that much of their work together with that of O'Donoghue *et al* (1997) and Thomas *et al* (1997) was motivated by a need to establish whether there were differences in outcomes between ethnic groups and, to a lesser extent, gender. As techniques improved, Goldstein (1997b:370) concluded that 'given what is known about differential school effectiveness it is not possible to provide simple...summaries which capture all of the important features of an institution'.

Goldstein's conclusion is of fundamental importance in that it effectively rejects the league table concept. He is supported by Ball's argument that 'schools cannot be represented by...single essentialising tags' (1997a:317). Ball said that 'schools are complex, contradictory, sometimes incoherent organisations' (ibid. p.317) and that the techniques that are intended to make schools more visible and accountable 'paradoxically encourage opacity and the manipulation of representations' (ibid. p.319).

Hence attention has more recently been directed at subjects and departments. Fitz-Gibbon (1992) provided evidence to support the conclusion that the school department was the desirable unit on which to focus attention not simply from an analytical viewpoint but because a department can actually use information on effectiveness. Talbert (1995:69) provided data from a large sample of secondary schools and showed significant differences between departments

which she related to 'sub-communities' of teachers. However, the issue of subject difficulty was raised and research by Fitz-Gibbon (1996:133) highlighted that some subjects were more difficult than others. This led to the development of 'relative ratings' that compare departmental performances within the school, taking into account subject difficulties and subject enrolments. Relative ratings have the value of showing how well pupils in a particular subject perform in comparison with their performance in other subjects taught in the school.

Differential effectiveness is an area where there has been relatively little research (Harris, Jamieson et al. 1995; Sammons, Thomas et al. 1997:10). Harris *et al* adopted a qualitative technique using departments that had previously been identified as 'effective' using QUASE data. Reasons for the lack of quantitative studies include concerns about the validity of available analytical techniques, that attempts to equate subjects in terms of difficulty could be problematical, and that different groups of pupils might have received different 'pedagogical treatment' (Schagen and Morrison 1999:3-4). Nevertheless, Saunders (1997:196-7) demonstrated that value-added measurements consistently reveal differential effectiveness for example girls outperforming boys, the negative effect of socio-ecomomic disadvantage, that (with the exception of pupils of Afro-Caribbean origin) ethnic minority groups tend to outperform white pupils, and that departmental differences exist within the same school. In terms of policy such findings are useful in pointing to where extra effort or resources are needed.

3.6 The special case of setting

This grouping of pupils according to ability in a particular subject is very well established in the UK although it is also a controversial one. 'Streaming', popular in the 1950s and 1960s, gave way to more 'mixed ability' teaching in the 1970s and 1980s, and this has in turn been substantially replaced by 'setting' in the 1990s (Boaler, Wiliam et al. 2000:631-2).

Simon (1991:307) linked the 'crucial issue' of streaming in the 1960s with the nature of the examination at 16; then the dual system of CSE and GCE 'O' level. He noted that the shift towards common syllabuses, the introduction of comprehensive schools, and evidence that streaming 'conditions' pupils to a level of response that this form of organisation sets for them resulted in a nationwide movement towards non-streaming or 'mixed ability'.

McCulloch (1998:147) commented that under the Conservative governments of the 1980s and 1990s the 'issue of social differentiation that had been largely concealed by the spread of comprehensive schools again became starkly evident'. He suggested a number of reasons for this, including the growing emphasis on differences both between and within schools that was emphasised in the White Paper *Choice and Diversity*. Boaler (1997c:576) postulated that a furtherance of moves away from mixed ability came after the introduction of the National Curriculum and that this happened partly because children are assessed and the level that they achieve is reported. It is also clear that 'mixed ability' was going out of fashion at this time and under the developing

'marketplace' of the 1980s parents were clearly expressing a preference for setting (Gewirtz, Ball et al. 1995:38-40).

The 1997 White Paper (DfEE 1997:11) claimed that 'the search for equality of opportunity in some cases became a tendency to uniformity'. The government's acceptance of inequality as a 'fact of life' to be acknowledged, rather than a problem to be overcome, was evident in the proposals that pupils *should* be taught in ability sets rather than mixed ability groups. The White Paper (DfEE 1997:38) strongly emphasised a commitment to setting:

> ...unless a school can demonstrate that it is getting better than expected results through a different approach, we do make the presumption that setting should be the norm in secondary schools.

The 'threat' in this statement is OFSTED but there is no suggestion as to what measurement would indicate 'better than expected' although clearly a value-added measure is implied.

Hence both direct and indirect influences of political pressures have had a clear impact upon student grouping policies in schools. However, Boaler (1997c:576) found little or no evidence to support the New Labour opinion that setting advances achievement. She suggested that schools are returning to policies of setting because they see it as a 'panacea to underachievement' (ibid. p.577). With a lack of evidence in mind she saw a need for new forms of research that will increase our understanding of the impact of student grouping policies upon student achievement.

A specific case of differential effectiveness occurs in situations where pupils are placed into sets according to ability in particular subjects. Using different techniques, Kilyon *et al* (1997:10) and Schagen and Morrison (1999:8) found that top sets over-perform and lower sets under-perform. In contrast, but using different methods, Boaler (1997c:577) reported that setting tends to produce some small increases in achievement for the students in the high sets gained at the expense of large losses for students in low sets. Looking at yet other studies, Ireson and Hallam (1999:347) found that generally the higher ability groups benefit, but there did not appear to be negative effects on the achievement levels of middle and low groups - or there was no effect. Negative effects for low ability groups were usually in terms of pupils' self-esteem and attitudes.

Hill (1996:3) and Ireson and Hallam (1999:353) reported that setting is another area in which few studies have been completed. However, in recent years there has been significant interest notably by the New Labour government and through the work of Boaler. Ireson was commissioned to report to the DfEE on ability grouping practices in secondary schools and guidance was subsequently sent to all secondary schools (Ireson 1999).

Boaler's (1997c:585) research, however, showed that a student's success in their set had relatively little to do with their ability, but a great deal to do with their personal preferences for learning pace and style. She found no evidence that setting raised achievement, but there was evidence (ibid. p.593) that setting diminished achievement for some students.

Boaler's work (1997d; 1997b; 1997c; 1997a; and Boaler, Wiliam *et al.* 2000) frequently points to fundamental questions about the process by which schools group students by 'ability' and subsequently how they are taught. These differences are many, varied, frequently subtle and difficult if not impossible to quantify. Class sizes (larger 'top' sets) and who is allocated to teach each set further complicate the picture. Ireson and Hallam (1999:348) for example, noted that many researchers have found that pupils tend to be labelled and stereotyped by teachers according to the group that they are in. Further, Nuttall *et al* (1989:774) introduced the notion of a 'compositional effect'. They hypothesised that, 'over and above the expected differences in performance attributable to differences between individuals attending each school, greater *concentrations* of underperforming groups will further depress performance (or vice versa)' (emphasis as in Nuttall *et al*). In the United States a study of 'tracking' (most closely associated with streaming in the UK) by Talbert (1995:79) found differences in goals, content, practices and expectations between tracked high school classes. In addition, teachers were assigned almost exclusively to either high-level or to low-level classes. All of this will inevitably lead to problems in determining the usefulness of any outcome measures and particularly in making comparisons between one class or teacher and another.

In the formation of groups, 'ability' is inconsistently defined. Some schools use measures of aptitude (for example cognitive ability test scores or YELLIS test scores), whilst others use attainment measures (for example Key Stage 3

subject scores or average KS3 scores). It is well known that the whole process of testing is not an exact science, as for example recently demonstrated by Williams and Ryan (2000). Behaviour is frequently a factor together with 'teacher recommendation' - and these are commonly superimposed on the objective measures. Boaler (1997c:590) found that social class appeared to influence achievement in setted lessons and this undoubtedly links with Ireson and Hallam's (1999:349) evidence that low ability groups tend to include disproportionate numbers of pupils of low socio-economic status, ethnic minorities, boys, and those born in the summer.

Boaler's (1997c:593) study revealed the individual nature of students' responses to setting. She considered that it was too simplistic to regard the effects of setting as universally good or bad for all students and reminded us not to overlook the complexity of the learning process for different individuals. However (1997c:591) she found, from a comparison of pupils in setted and mixed ability situations, that 'open', 'progressive' models of teaching in a school with mixed ability groups partly explained better results.

3.7 Teacher effectiveness

The question remains as to whether value-added techniques can be used in the assessment of teacher effectiveness. From a policy point of view this is highly desirable and in keeping with the notions of accountability, Threshold Assessments and Performance Management. It could be cost-effective and efficient if teaching quality could be quantified and the data made available to

managers. Measuring teacher effectiveness is also desirable to complement the judgements made by OFSTED since, for example, inspections only effectively give a snapshot of a teacher's performance during an inspection week. A value-added measure could be both more reliable and have greater validity in that it covers a longer time period of teaching.

From a practical point of view there are issues of a statistical nature to be overcome - small amounts of data, 'ownership' of the results (for example where classes are shared), the contributions made by former teachers and the plethora of other factors that get in the way of making valid comparisons between individuals. Certain teachers may feel excluded because the nature of their work means that they do not generate the sort of 'output' data that the majority do, for example teachers of pupils with Special Educational Needs. There would also be ethical considerations - just as it has been shown to be unhelpful to sum up a school in a single statistic, so it would be morally wrong to do the same for individual teachers. However, it must be remembered that, for example, Threshold applications include a measure of pupil progress as only one out of five categories in which teachers need to provide evidence.

Until relatively recently classroom practices have remained remarkably stable over time. Evidence has suggested that policy that proceeds down without a consideration of the lives of teachers is unlikely to completely succeed - Osborn referred to this as 'creative mediation' (Osborn, Croll et al. 1997:52) on the part of teachers. Reynolds, however, argued that there has not been enough pressure on teachers:

> We have, by intervening with the school level rather than with the learning level, been 'pulling levers' that have small effects on their own and which may not have generated any 'ripple through' to affect the key level of the classroom (Reynolds 1999c:73).

Yet contemporary evidence, for example from Helsby, showed that:

> ...recent educational reforms have enjoyed a notable degree of success not only in moving the goalposts but in revising the basic rules of the game (1999).

The government's interventions in pedagogy still have some way to go. In a commentary on the revised National Curriculum, Elliott (2000:254) stated that it is 'over-optimistic to assume that significant pedagogical change will automatically follow' with a more flexible curriculum - indeed he questioned whether the curriculum is flexible enough. Harris (2000:5) suggested that future school improvement work needs to provide evidence of impact upon student performance and also needs to be more closely matched to the needs of different types of schools (ibid. p.8). Thus her solution lies in a focus on *'what works* in practice, rather than *what fits* in terms of political expediency' (ibid. p.1; emphasis in original).

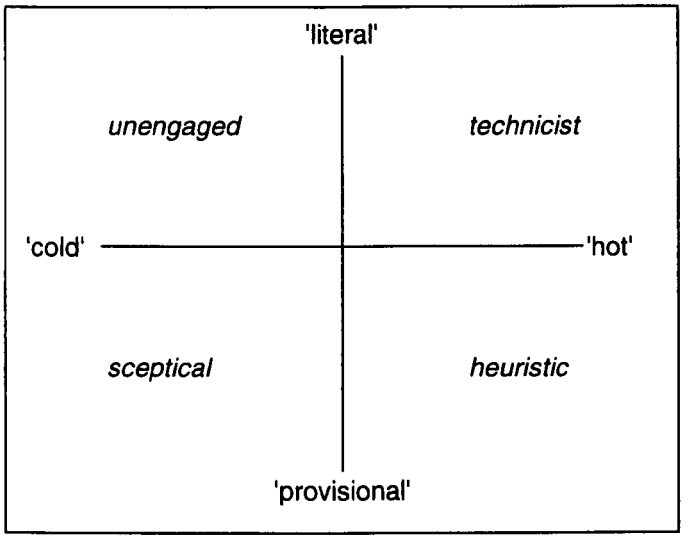
Traditionally, it is 'what works' that has been the 'measure' of an effective teacher. Most research on effective teaching has been of a qualitative nature frequently involving lesson observations. Many studies have sought to identify effective teaching behaviours through 'process-product' enquiries whereby the researcher has sought to correlate teaching behaviours (process) with student achievement (product) (Harris 1998:171). Other studies reviewed by Harris seek, for example, to identify effective teaching skills, styles and models. The result is a list of the characteristics of the 'effective teacher' (for example see

Scheerens 1992; Cullingford 1995) which now form a checklist for OFSTED inspection teams (OFSTED 1995).

In his exhortation of teaching as a science, Reynolds (1999a:13) produced a similar list of characteristics - good teachers, for example, have high expectations, challenge pupils intellectually and use time well. However, he felt that this should be codified and turned into a body of 'science' and 'with this body of science...come instruments to measure what teachers do'. Thus Reynolds wants to codify best practice and develop performance indicators to use as benchmarks. This has been produced (and validated by Reynolds) in the form of the Hay McBer Report.

An imperative in the search for ways in which value-added can be used in the assessment of teacher effectiveness is to examine the ways in which teachers currently use such data. Saunders and Rudd (1999:20) found a 'spectrum of attitudes' in their exploration of the use that middle managers in schools made of value-added data. They considered a model to describe the different approaches to performance data, which was subsequently elaborated by Saunders (2000:241). She considered the technical, conceptual and ethical issues that schools encounter at senior and middle management levels and developed the model to represent different emotional and intellectual stances towards the use of performance data. This is shown in figure 2 below.

Figure 2 Saunders' (2000) model to represent different emotional and intellectual stances towards the use of performance data



In this model 'cold' and 'hot' represent emotional attitudes (degree of enthusiasm) towards performance data. 'Literal' and 'provisional' indicate intellectual stances, that is, the degree to which teachers rely on data as a manifestation of pupils' ability in a given subject. The four quadrants thus formed and named describe different combinations of intellectual and emotional responses to the use of data. An important aspect of the current study is to test the usefulness of this model in the analysis of teacher interview data and to assess whether it helps in developing an understanding of what works in the use of value-added in classroom practice.

3.8 Concluding comments

It must be concluded that value-added has entered the 21st century in a problematical situation. Government, until Gillian Shephard's interest in 1994,

effectively shunned the concept and proposals to establish a National Value-added System have effectively been placed on a 'back burner'. This is evidenced by the distinctly low-key responses to value-added research, notably the Value-Added National Project, together with the introduction of the 'alternative' systems of the PANDA and Autumn Package. Saunders (1999:253) has commented that the key players in value-added research (Fitz-Gibbon, Tymms, Thomas and Goldstein for example) have now argued against public accountability use of value-added analyses. Goldstein (2001:437) has recently stated that the attempt to associate progress for a pupil with a single teacher is not only divisive, it is also likely to be misleading. Pockets of distinct antagonism to the concept remain, for example, in a new publication by Donnelly and Jenkins (2001:162) words such as 'grotesquely' and 'offensive' are used about value-added.

Saunders (1999:253) believed that although the principle of value-added has been understood, the more rigorous the value-added analysis becomes the less simple and straightforward the interpretation becomes. In her words:

> ...the value-added task began by appearing to promise better information for public consumption, but instead it turned out to demonstrate that 'better information' and 'public consumption' are incompatible.

However, value-added has a distinct and growing role in school management. Schools value the information that they receive from systems such as YELLIS and will often use it to inform raising-achievement initiatives or to 'mitigate the message given by their raw results' (Saunders 1999:252) - data is, for example, frequently passed to OFSTED during inspections. Increasingly the base-line

data is used in target-setting with the value-added feedback being used in the assessment of the extent to which targets were met.

It is also in the area of Performance Management that a problematical future for value-added lies and the Government's advice on using such data is 'at best confusing and at worst misleading' (Goldstein 2001:435). In the first round of the Threshold Assessment process most teachers who applied were successful, perhaps in some cases as a result of the 'light touch' that was applied by assessors. With the second round in the Autumn of 2001 along with the completion of the first cycle of Performance Management, it remains to be seen how much value-added measures will feature in the assessment of teacher effectiveness.

Chapter 4 From policy to classroom practice at Dalestone School.

I thought that we could do better if we could put our finger on it - well for certain - was it pie in the sky? Was my instinct wrong and we couldn't do any better? I didn't feel that I was wrong. I believed that there was more to come, especially from lower down the notional ability range. (Former Headteacher of Dalestone School talking about why he took the school into value-added. 16-08-00, Interview record).

NOTE: Where there are references to residuals or test results for a particular year, the year given is when that particular cohort took their GCSE examinations. Thus the 1996 cohort were in Year 11 in 1996 and sat their GCSEs that summer. They would have taken the YELLIS test in the autumn of their Year 10, that is, in 1994.

4.1 Introduction

In this chapter an historical account is presented of the development of value-added policy and practice at Dalestone School. This is set within the context of local and national developments.

The school was introduced to value-added partly on the instinct of the Headteacher and partly in response to a failed local authority initiative. A subscription to YELLIS resulted but in the early years the data was treated with the utmost caution. A new Headteacher was appointed shortly after the formation of a new Unitary Authority and both adopted a data-rich 'accountability' philosophy in keeping with the trend adopted by the government. Initially there was no place for YELLIS in a regime where a new Deputy Head adopted a confrontational style in holding staff to account for their targets and results.

The use of YELLIS however, continued to be popular with teachers. Its consistent use in target-setting and analysis of results has resulted in a new policy-shift towards this value-added tool. Most recently it has been used extensively by teachers in reflection on their practice linked to applications for Threshold Assessment.

4.2 The School Context

Dalestone School is an 11 - 18 co-educational comprehensive school in a Northern Unitary Authority which shall be referred to as Chalkshire in this study. The school has over 1,220 pupils on roll and is poised to grow to 1,450 pupils over the next few years as the local population increases. Dalestone's mission statement emphasises 'equal worth' and this is significantly at the heart of the school ethos. Chalkshire was formed in 1996 following the Conservative Government policy to create Unitary Authorities - its predecessor authority, Rockside, was split into four.

The school serves a part-rural, part-commuter community in an area of approximately 120 square miles. It is centred on a small market town (population around 8,000). The school has had three Headteachers since opening in 1953. The second and longest serving Head, who retired in 1997, was instrumental in the introduction of value-added at the school.

The March 2000 PANDA report (OFSTED/DfEE/QCA 2000, School records) records the following profile of Dalestone school:

- ◊ 4.9% of pupils are eligible for free school meals compared to a national average of 18.1%.
- ◊ The number of pupils with Special Educational Needs (7.5%) is well below the national average (18.7%) but the number with Statements of Special Educational Need is broadly in line with the national average.
- ◊ Very few pupils are from ethnic minority groups. Only one pupil has English as an additional language.
- ◊ The attendance rate and the rate of unauthorised absence are broadly in line with national averages.

PANDA information about the school's context is based on the, rather dated, 1991 Census of Population. This showed around 12% of adults with higher education, around 35% of children in 'high social class households' and very few (<5%) of 'children in overcrowded households'. It is anticipated that the 2001 Census will reflect the immigration of a significant number of families to newly-built homes in the area with adults taking relatively well-paid jobs in, for example, the two cities which are within easy travelling distance.

In the analysis of evidence about value-added policy and practice at Dalestone School it was clear that there have been three phases of development. The first phase, from 1993 to 1995, was characterized by caution and a concern for 'correctness' in the use of the new source of performance data. From 1996 to 1999 the school experienced significant change following an OFSTED inspection, a change in leadership and having to respond to the new (national) 'target-setting' agenda. The last two years have been dominated by the movement of pupil progress measures into the arena of teacher effectiveness and in particular Threshold Assessment.

4.3 1993 to 1995: A cautious approach to value-added

Three factors were significant in setting Dalestone School on the value-added path. One was the 'raising achievement' agenda that had emerged following the phased introduction of the National Curriculum in the early 1990s. The second was concern over recruitment into the sixth form. The third was a paper that had been produced by a Rockside LEA statistician in January 1993 entitled *Value-Added Analysis - Performance at GCSE.* (Woodward 1993, School records)

As the quotation at the start of this chapter suggests, the 'raising achievement' agenda at Dalestone School has its roots the Headteacher's 'gut' feeling. However, he was the sort of person who wanted hard evidence before he would act. Papers that the Head had read convinced him that there was enough evidence to say that some schools do better than others having similar characteristics. Coincidentally, he also read an article in the *Times Educational Supplement* by Professor Fitz-Gibbon (then at Newcastle University) about the start of the YELLIS project which 'intrigued' him for three reasons. Firstly, the article suggested that YELLIS would provide the sort of evidence that he was seeking. Being a 'North-Easterner' and having worked for the Schools' Council in the North East in the 1970s, he felt that he had a sense of the sort of schools that were in the project at the start. Some he had personal knowledge of and he felt that they were not dissimilar to Dalestone School. Finally, he knew Newcastle University and the quality of the work that it produced (he had studied there).

Although Local Management of Schools (LMS) was still young, Dalestone went into a budget deficit of £55,000 in 1992-3. This was essentially the result of an over-optimistic forecast of Sixth Form numbers such that the LEA 'clawed back' money that had been allocated the previous year. The Head was anxious about recruitment as he recalled later:

> The Sixth Form was too small for comfort at the time and too extravagant. If we could increase the staying-on rate by knowing who was capable of work at a higher level we could potentially make more economic groups post-16. We did put on more vocational courses and YELLIS was influential in that. (Former Headteacher, 16-08-00, Interview record)

A group of Rockside Headteachers formed a Post-Compulsory Forum. This was an unofficial group that was chiefly concerned about recruitment, particularly because of competition from other institutions. In Dalestone's case there were three main competitors - the local independent school and two Further Education colleges. The Forum was eventually 'adopted' by Rockside and this paved the way for the future funding of value-added schemes.

Rockside's *Value-Added Analysis - Performance at GCSE* suggested that Dalestone School was performing rather poorly in value-added terms. Governors and senior staff at the school were angered by the publication for a number of reasons - it had been published across the County, the school was readily identifiable and there had been no consultation with the school about the data that it contained. The school's response was initially to complain as did a number of others. Subsequently, the school established a Working Group to examine the question as to whether it was indeed underperforming.

> I didn't think that we were doing badly although the Rockside report was a bit of a sting. I wasn't happy about being below the line. I didn't think that we were significantly underachieving - but the suggestion was strong

> enough for me to say 'let's have a look at it' (former Headteacher, 16-08-00, Interview record).

Value-Added Analysis - Performance at GCSE placed the notion of value-added firmly on the agenda but concern about the quality of the report encouraged the school to search for a 'fairer way'. In September 1993 Dalestone joined YELLIS and was one of only 49 schools in the project at the time. The following year 242 schools were signed up indicating the rapid growth in interest that schools were showing in value-added. Rockside LEA offered to fund a few schools' involvement in YELLIS. These were selected from the Post-Compulsory Forum and included Dalestone. Rockside supported the subscription partly to move away from the embarrassment of the January paper. The LEA also wanted more information about baseline testing and in particular there was a desire to evaluate the YELLIS test. Also, the LEA was able to tap into TVEI (Technical and Vocational Education Initiative) funding to pay for the pilot. Later, the LEA supported the school in joining the ALIS project. Eventually the school had to pay the subscription initially through the staff development budget and later from the delegated budget.

The Rockside analysis had explored the correlation between performance and a number of socio-economic characteristics of the 60 LEA schools. In addition to using the usual 'percentage free school meals' the analysis was conducted using:

- ◊ Clothing grants per 100 on roll
- ◊ Percentage male unemployment in the nominal catchment area
- ◊ An amalgamation of Census data (1981) reflecting socio-economic need in the nominal catchment area. (Woodward 1993, School records)

Schools were ranked on each parameter and a final socio-economic ranking was produced - Dalestone was ranked 9th. Performance parameters (such as the percentage of an age group obtaining 5 or more A*-C grades at GCSE) were also ranked. The rank scores were plotted on a scattergram which also displayed a regression line describing a perfect correlation together with lines which were effectively confidence limits. In terms of GCSE performance, Dalestone was placed close to the lower confidence limit. For Mathematics, the school was shown as seriously underperforming.

Dalestone's 'Raising Achievement' working group consisted of 11 Heads of Department most of whom were long-serving and experienced teachers. It was not until the summer term 1994 that they made progress and held a number of meetings prior to publishing a report. By now the 1991 Census data were available and this showed that Dalestone had risen to third place in the LEA socio-economic ranking. The group demonstrated that GCSE results had improved steadily and the notion that Dalestone was an underachieving school was seriously challenged. The statistical model was shown to be flawed in that it was now impossible for the school to become overachieving! The Head of Mathematics found further faults in the statistical methodology. The methods used in ranking and in producing so-called 'regression' lines were dubious. In the Value-Added National Project Final Report four years later, Fitz-Gibbon commented that 'rank orders are inherently undesirable because they fail to convey the *size* of the differences' (Fitz-Gibbon 1997: section 4.8, emphasis in original).

A meeting of governors, the Senior Management Team (SMT) and the Heads of Department was convened and the LEA officer who had written the report accepted an invitation to attend. She admitted to the failings of the statistical model which was subsequently shelved. She stated that the LEA still intended to look for links between socio-economic factors and pupil achievement. Concerns over the use of socio-economic data prompted the Working Group's main recommendation, that the school should collect baseline achievement data and continue with the value-added (YELLIS) work that had just started (Raising Achievement Working Party 1994, School records).

The report contained several other comments and recommendations that mainly focused on curriculum and management issues. Factors such as reductions in staffing levels, financial constraints, the school calendar and the timing of the school day were presented as factors that had affected achievement. There were calls for timetable arrangements that would permit more setting of pupils, and more support for able pupils. Significantly one brief paragraph entitled 'Good teaching can turn tables' was included about teacher effectiveness:

> In compiling this report we have taken good professional practice for granted, but it may be worth quoting a report by Ted Wragg (*The Observer* 19th June, 1994) comparing classrooms in the same school.
>
> > "The differences were clear. The two effective teachers were well organised, clear and consistent in their classroom rules, set interesting and challenging work, put out positive messages, and were quick to recognise success or help those who were struggling. The third was the opposite. Her classroom rules were ambiguous and inconsistently enforced, expectations were low so little demand was made, and she simply whinged at the relentless misbehaviour that ensued from endemic boredom".
>
> (Raising Achievement Working Party 1994, School records).

It was clear that the Heads of Department were aware of the significance of teacher effectiveness but they wished to raise the issue in a subtle way particularly since their report was to have a wide circulation.

At this time the SMT's policy was that YELLIS would be used as one of the strategies to meet the 'raising achievement' school development plan objective. This was very much in keeping with the ethos of the school that stresses concern for the individual. YELLIS data would be used to monitor, plan for, and mentor individual pupils. The Head was particularly active in promoting YELLIS to the SMT and Governors. He delegated the management of the project to the Senior Teacher with responsibility for assessment, recording and reporting.

There was no suggestion from the SMT that YELLIS was to be used to measure the work of teachers and little thought was given to the potential use of value-added in measuring the performance of the school. The Senior Teacher attended various meetings and subsequently recounted concerns about what were considered as negative uses of value-added data. In particular he was concerned about any mass publication of data, setting up one subject against another and the potential to hold staff accountable for the residuals obtained by their classes (Interview with former Senior Teacher, 24-06-00, Interview record). The CEM Centre still urges caution about issues of publication and confidentiality in their *YELLIS Handbook* (Curriculum Evaluation and Management Centre 1999:30). Thus the Head decreed that initially only governors and managers would be privy to YELLIS data although clearly other

staff would be aware that, for example, the YELLIS test was taking place in the Autumn.

> It could be quite threatening to staff really. If it was going to expose that we could do better then it perhaps was going to ask questions about which parts of us could do better - could we all do better? We had two star departments [English and Mathematics] which were generously treated for curriculum time and had opportunities for setting which others did not.... I was also conscious of the variability in our intake - we would need a way to iron out the kinks and avoid painful inquests if there was a bit of a dip (former Headteacher 16-08-00, Interview record).

It was not until January 1995 that the Head mentioned YELLIS and ALIS to the whole teaching staff during a training day (former Headteacher 1995, School records). His remarks prompted expressions of concern from teachers about the lack of information about these projects that the school was involved in. Realising the mistake of keeping the projects so secret, the Head invited staff to a talk that had been arranged for the governors' curriculum committee the following week. He also circulated some literature.

With only 5 days' notice, few teachers other than the SMT attended. My notes from 10th January, 1995 show that the speaker from the CEM centre outlined the statistical background to YELLIS and ALIS and made five key points. His message effectively summarised the school's value-added policy although this was never formally written down. Value-added data would:

- ◊ provide quality evidence to support professional judgements.
- ◊ be used to inform 'alterable school policies' such as timetable and resource allocations.
- ◊ be used to inform 'alterable classroom activities' - i.e. teaching and learning styles.
- ◊ be used to provide guidance for students and in particular to build confidence.
- ◊ be used to identify the 'outliers' and marginal students (Rogers 1995, School records).

When interviewed, the Senior Teacher remarked:

> We tried to operate in as gentle a way as possible. We were non-threatening. YELLIS was seen very much as a tool for management and governors. At that time we could see that YELLIS had the potential for what is now called 'target-setting' and 'Performance Management' but these were seen as threatening and we were not prepared to entertain this....The Head was a gentle man - he would never want to hurt people; he was a people person. In the long run he saw more benefits from using YELLIS in a positive way for the benefit of children. He prevailed upon the governors to keep the pace of implementation gentle. I was a teacher governor - the governors then were a very united body and very sensitive (former Senior Teacher, 24-06-00, Interview record).

This sensitivity resulted in many long debates amongst the SMT and governors about the use of YELLIS. In particular the consensus was that the publication of the first set of residuals in 1994 should be restricted. Parents were not informed and Heads of Department only received data for their own subjects.

The Head of Mathematics, also a teacher governor, was the only member of staff outside of the SMT to be significantly involved in YELLIS. Having been 'stung' by the Rockside analysis of 1993 he adopted a cautious approach and particularly advocated that any one year's results did not matter as much as overall trends. Some training (in the form of a number of after-school meetings) was provided for Heads of Department but more time was spent on the philosophy and methodology of value-added than actually looking at the first set of residuals.

By the summer of 1995 there was a realisation that the school should be getting a good set of GCSE results from the Year 11 cohort. YELLIS categorised 41% of pupils as being in the most able 'band A' (no subsequent year has had such a high proportion). This cohort had taken the YELLIS test at the 'proper' time in Year 10 whereas the previous group had taken it in Year 11. Thus there was an

air of confidence about the validity of the data, an expectation that residuals would be better and also some familiarity amongst Heads of Department about the whole concept of value-added. At the same time the school was expecting an announcement about when OFSTED would visit. It was felt that the school would be criticised for having data and not making wider use of it. In this context a full training day for the whole staff about value-added was organised for late October 1995 - again to be led by the CEM centre.

The day was again carefully planned to be non-threatening. The representative of the CEM centre was well aware of the school's anxieties. His letter to the Senior Teacher confirming arrangements includes:

> I enclose a draft programme for the 20th, which I hope bears some relation to what we said! Please alter it at will, and just tell me on the day...we must certainly get it right (Elsom, 1995, School records).

By October, the 1995 data was available and subject teams discussed their results in workshop sessions during the training day. The GCSE results had been good - the 5+ A*-C grades were up by 6% to 51% and residuals in most subjects were also better than 1994. Hence the training day had a very positive edge.

One drawback was that the bulk of the data was being used in a retrospective fashion. Heads of Department and subject teachers looked at results and made statements about how well or poorly particular students had performed. Reasons for each 'case' were proposed but the level of discussion was superficial. The argument that we only had data covering two years and therefore could not discuss trends held sway.

The YELLIS test results for the 1996 cohort were available and could have been used in a proactive way but these were kept from staff. Although 'policy' intended that YELLIS was to be used to raise achievement, paradoxically there was a sense that we should not intervene. The force of the argument was that if we pressed to add value at Key Stage 4 then we might distort the picture that would eventually emerge post-16. Fitz-Gibbon (1997: section 1.3) states that 'much is yet to be learned about the extent to which the value-added indicator can be altered by the efforts of schools'. Indeed, the Value Added National Project Final Report expressed concern about the temptation to 'fiddle' data in a 'high stakes' system or the consequences for one Key Stage of good results in the previous Key Stage (Fitz-Gibbon 1997: section 2.4). This was appreciated as we entered YELLIS, thus there was no strong suggestion that staff could make use of YELLIS test results and intervene and actively work to help a pupil perform better. Another reason for this links with the school 'equal worth' ethos - it would have been easy to select those pupils on the GCSE grades C/D border for extra attention but this would have been in contradiction to the school's mission statement.

In the first two years most use was made of the YELLIS 'school profile'. As part of the YELLIS test, students answered questions about, for example, their aspirations and use of libraries. The YELLIS feedback included a profile comparing the school's pupils with the whole YELLIS cohort in terms of 'cultural capital' (use of books etc), career choice and job aims. This was found particularly useful in planning for post-16 recruitment. Also additional resources

were put towards the library since pupils had been significantly negative about it.

The only area where YELLIS data were used in relation to concerns about teaching was with the Modern Languages Department. Here, YELLIS data for 1994 in all three languages recorded significantly negative residuals, confirming serious underperformance. The Head asked me, as the link member of the SMT, to investigate and take the necessary action.

Staffing issues were at the root of the Department's problems - exacerbated by budget constraints (Rogers 1997:15). Leavers had been replaced by several part-time staff - or not at all. The employment of Foreign Language Assistants also ceased. There were also difficulties in motivating a sizeable minority of pupils, obsolete and depleted stocks of text books, a condemned language laboratory and staff teaching in mobile classrooms. Historically, changes to salary scales and staff changes had led to an erosion of salary points resulting in the Department having a relatively low status and its members feeling undervalued.

Progress was made with the restoration of six full-time staff, pupil motivation and in department documentation. Some use was made of the 1994 residuals, and the 1995 and 1996 YELLIS test results, most significantly resulting in a revised approach to setting arrangements. The progress was reflected in improved residuals for 1995 and 1996 in two of the languages. By 1996 the OFSTED report recorded standards in GCSE as being about the national

average, the quality of learning as 'sound or better in the majority of lessons', and most of the teaching as of a 'high quality' (Baxendale 1996: paragraphs 59-62).

Caution, wide consultation and beginning with small amounts of data were all advocated by the Value Added National Project Final Report (Fitz-Gibbon 1997: sections 2.6 and 4.1). This had been the policy at Dalestone over the 1993 - 1995 period. Such a cautious approach seemed to have paid dividends in terms of acceptance by the teaching staff but in terms of using value-added to raise achievement, progress had been patchy and generally slow. Although knowledge about value-added had increased, the proactive use of YELLIS data was very limited.

4.4 1996 - 1999: New agendas

The first OFSTED inspection was in January 1996. This was the first inspection since a Rockside Advisory Team inspection in 1990. The OFSTED inspection team was complimentary about the school's efforts in gathering information about pupil attainment. In particular OFSTED praised the use of data in setting arrangements in Mathematics where 'subsequent diagnostic work is effective in helping to produce good GCSE results, especially at the highest A* and A grades' (Baxendale 1996: paragraph 13).

However, the Inspection Report was particularly critical about aspects of the management of the school and this included the management of value-added:

> Participation in widely-used diagnostic assessments of potential in GCSE and A-level, conducted at Key Stage 4 and in the sixth form provide substantial data which needs enhanced management to be fully effective (ibid. paragraph 13).

The OFSTED inspection inadvertently prompted a hiatus in the school's development in that it pointed to the Headteacher's anticipated retirement. The top two key issues related to a need to reform the management of the school. However, the post-OFSTED Action Plan postponed change until a new Headteacher was appointed.

Political issues characterise this period. In April 1996 the new unitary authority of Chalkside was established. Being a 'hung' Council the process of appointing LEA officers was slow and this authority was worse off than Rockside in financial terms per head of population. Relationships between Dalestone's governing body and the new LEA were soured at this early stage as a result of a major disagreement over the appointment of the new Head who took up his post at Easter 1997. In May 1997 the Conservative Party lost the General Election and New Labour came to power.

The new authority showed great enthusiasm for the collection and analysis of assessment data and a Research, Statistics and Quality Assurance (RSQA) team was set up. The 'fresh start' and being a smaller authority (18 secondary schools) provided the opportunity for Chalkshire to quickly establish baseline testing in all schools. Thus all secondary schools now subscribe to the NFER Cognitive Ability Tests (CATs) and these data are used alongside other assessments such as Key Stage tests and GCSE results. Chalkshire clearly indicated its intention to meet the demands of *Excellence in Schools* (DfEE

1997:26) and could be described as being influential as a 'driver' of school improvement in contrast to Reynolds' general view of LEAs (Reynolds 1999c:75).

In May 1996 the DfEE and OFSTED produced a report *Setting Targets to Raise Standards: a survey of good practice.* It said that 'inspection shows that internal review and target setting are often the weakest parts of the school's planning cycle' (1996:paragraph 1). It contained an example of a school using YELLIS but commented:

> While the GCSE indicators are informative, one shortcoming is that the indicator is based on an attainment test taken at the end of Year 10 which allows the school little time to provide much formative help (ibid. para. 43).

This statement was uncomplimentary about YELLIS and it was also inaccurate in two ways. First, the YELLIS test is an aptitude test, not an attainment test. Second, the test need not be left to the end of Year 10 - it can be taken in the first term of that year. Although probably not influenced by this report, the LEA was not willing to financially support schools' use of YELLIS and the Senior Teacher believed that some officers also adopted a negative view of value-added (former Senior Teacher 24-06-00, Interview record).

The Senior Assistant Director of Education proposed a framework of LEA targets in November 1996 (White 1996, School records). A series of meetings followed in the spring of 1997 (White 1997a, School records). The target-setting theme was in anticipation of legislation (the 1997 Education Act) and the LEA was keen to develop an analytical approach.

A Chalkshire document, *Setting Targets to Raise Achievement,* was published in June 1997 and this established their approach (White 1997b). Although after the General Election, it had much in common with the framework suggested the previous November. A 'bottom up' model was prescribed where whole school targets are derived from individual subject targets, which, in turn, are based on targets agreed with individual pupils. The LEA also detailed the evidence on which subject target-setting at Key Stage 4 should be based. This included:

- ◊ recent profiles and trends of attainment in the subject at GCSE within the school.
- ◊ comparisons of profiles of attainment and trends with other subjects within the school.
- ◊ consideration of relative progress from Key Stage 3 National Curriculum tests and/or CAT scores in the subject within the school and, if the data are available, in comparison with other schools in the LEA.
- ◊ consideration of the prior attainment of the students in terms of evidence from Key Stage 3 National Curriculum assessment results, CAT scores and other standardised tests and evidence of what students with similar prior attainments have achieved in the school in the past. (White 1997b, School records)

The amount of data analysis expected was thus considerable. This was to be multiplied by similar requirements at Key Stage 3 and post-16. The LEA quickly repulsed the inevitable negative reaction from some schools. The RSQA team visited schools, offered training and support and undertook some analytical work. GCSE data for the LEA (derived from the National Consortium for Examination Results) was made available on computer disks and software to help with analysis and setting targets post-16 was developed. Ultimately however middle managers would have the key role:

> The acquisition of the data, whether from the LEA or from systems such as YELLIS will not of itself lead to the establishment of a target setting culture in the school. Nor will such a culture develop if the analyses are undertaken by a senior member of staff and not shared with heads of department. Ideally, heads of department need to be supported to undertake the analyses for themselves so that, by undertaking the process, they can come to appreciate possible issues and what smart

> subject specific targets might be (White 1997b. paragraph 4.8, School records).

This is the only reference (albeit an indirect one) to value-added. The Authority had decided not to support measures of value-added, including YELLIS, preferring to promote measures of 'relative progress' as an alternative. 'Relative progress' differs from value-added as defined in this study in that it measures the difference between a baseline assessment and an actual result rather than the difference between the result obtained and the result predicted from a baseline assessment. Further, the 'relative progress' measures would be restricted to individual schools or would use aggregated data to make comparisons between schools. This is fundamentally different to the use of pupil-level data in a national cohort as is the case with YELLIS.

The LEA was forceful in the promotion of its Framework. The final paragraph of the June 1997 document illustrates this hard edge:

> There are concerns that such an approach can lead to an over-preoccupation with "hard" targets which, in turn, may lead to other areas of student achievement and education provision being devalued. This is not the case. This approach to target setting can and should be applied to any area of development. Whatever is valued can and must be evaluated and targets can be set to provide focus and direction towards achieving particular goals (ibid. paragraph 5.5).

In his first term in office the new Head of Dalestone School took the opportunity to adopt this strong 'accountability' line following the resignation of the Senior Teacher (Assessment, Recording and Reporting) and the first Deputy Head. New appointments permitted a rapid reorganisation of the management structure (an OFSTED key issue) and the establishment of a new regime. Whereas the 'long-standing philosophy and ethos [had been] founded on

integrity and responsibility' (Baxendale 1996: paragraph 124) the new philosophy was very clearly founded on accountability. This shift in emphasis clearly reflects the national agenda, but internally the new challenges to the professionalism of middle managers created significant tension as illustrated by a comment made by one long-serving teacher referring to the new Deputy Headteacher :

> I put Rikki's target down as an E. He's never liked Science and we've always had a battle when it comes to homework. I also know that he found maths hard but his English was OK. He only got a level 4 Science Key Stage 3. But, his CAT score was over 100 and she said that made my target too low - she wanted me to make it a C. I said no, but she wasn't satisfied with my professional judgement. That's just insulting and shows no real grasp of stats - just 'cos the average child with a CAT of 100 gets a C doesn't mean that they all will - some will be below. (Teacher N, male, Science)

In the Autumn Term of 1997, Chalkshire produced an analysis of examination results subject by subject and school by school (White 1997c, School records). Schools were named, much to the chagrin of Headteachers who remonstrated with the LEA but then became quiet. The LEA produced a further document in December, *Key Indicators* (Curriculum and Quality Development Team 1997, School records), this being an LEA equivalent of a PANDA but containing more information on pupil characteristics, staffing and analyses of spending. The White Paper, *Excellence in Schools* stressed the importance of schools looking critically at their pupils' attainment data as an essential first step towards improving performance. Analysing results and setting targets were highlighted as essential parts of a school improvement strategy which also included wider issues such as the development of Information Technology.

The New Labour Government was quick to set up a Standards and

Effectiveness Unit which produced two followup booklets to *Setting Targets to Raise Standards: a survey of good practice.* Both *From Targets to Action* (1997) and *Setting Targets for Pupil Improvement: Guidance for Governors* (1997) provided detailed guidance on target-setting, benchmarking and analysis of data but contained nothing about value-added. In response at Dalestone, the new senior staff were busy setting up a Central Curriculum Record (CCR) to meet a key Development Plan target (Dalestone School 1998, School records). This database now contains pupil assessment data (KS2 and 3 results, CAT scores, targets, GCSE results etc) together with any information about Special Educational Needs.

The GCSE results for the 1998 cohort were almost the best ever, with 58% obtaining 5 or more A*-C grades (59% in 2001). However, the YELLIS analysis was comparatively poor - only History, Mathematics and Double Award Science had positive residuals. In her commentary on the YELLIS analysis, the Deputy Head (Curriculum) (Deputy Head 1998a, School records) wrote:

> This is not the most positive analysis of our exam data, made all the more uncomfortable by the delight we all shared in the best absolute data that the school had generated.

Clearly uncomfortable herself about the analysis, and also caught in the government and LEA anti-value-added spin, she goes on to say:

> Some schools in the LEA are now opting out of YELLIS and ALIS given the increase in data available from other sources and the new format of the performance tables....If colleagues feel strongly that the YELLIS analysis is unreliable, we ought to seriously rethink our membership of the scheme in future years (ibid).

Yet she is clearly not convinced about that strategy since she concludes the document by noting that underachievement may be concentrated in particular

pupils or groups and that;

> Careful reflection on the YELLIS results might make the picture clearer and help inform strategies to raise achievement in the future (ibid).

As value-added was having a difficult time in school, so there were problems at a national level. Having had time to consider the contents of the Value-Added National Project (final report published in February 1997) the New Labour government developed a Value-Added Pilot Project and planned to publish a 'progress index' showing improvement from Key Stage 3 to GCSE. Within weeks of publication of the 1998 Performance tables the 'index' (a scale from A-E) was dropped and only those schools making 'good' progress were highlighted by a tick in the tables. Cassidy (1998:4) quotes Professor Harvey Goldstein's criticism about 'crude and simple mathematics' used in the pilot and also reports that protests from schools caused the index to be dropped. The 1999 tables did not contain any value-added measures. Cassidy's (1999:4) report highlighted several statistical problems that the government had not yet overcome - 'no "value-added" measure is now expected to be published until at least 2003'.

However, a key feature of the YELLIS system is the test from which predicted grades and ability bands are derived, and the opportunity to use 'chances graphs'. So, although value-added was on the 'back burner' in a national sense, use of the system in relation to target-setting at school level became important - particularly at departmental level (Rogers 1999, School records).

The first target-setting exercise 'Targets 2000' proved to be a tremendous

challenge for the school. The LEA required the school's targets by 16th October, 1998 and the school decided to follow the time-consuming, 'bottom-up' approach. Thus although at the start of the Autumn Term staff were briefed about the process, work on setting targets was delayed by the analysis of the 1998 results and by problems in the production of teaching group lists. The prior attainment data was available by the end of September but the results of the YELLIS test were not.

A training day in October 1998 was used to follow up the process and to refine the targets. Teachers felt pressured to raise their targets by the internal publication of all subject- and pupil-level targets, the accusation that targets did not include an element of challenge, that there should be no non-entries, and that there were too few A* and A grade targets (Deputy Head 1998b, School records). Consequently, six subjects raised their targets and by the beginning of November the 'bottom-up' target had reached 59% 5+A*-C grades.

Many staff were unhappy about the mass of data to work through and the short timescale (Rogers 1998, School records). There were also complaints from staff having to set targets for pupils that they had not taught before the start of this term (Rogers 1998, School records). The Deputy Head (Curriculum) was dismissive of this, suggesting that there might be a problem with department record-keeping at Key Stage 3 and that SMT members might 'nudge departments who haven't got the appropriate systems in place in the right direction' (Deputy Head 1998c, School records). Yet departments such as Business Studies, with no input at Key Stage 3 considered this an unfair

statement. For them the YELLIS test scores provided a new baseline in which they had more confidence (Rogers 1998, interview record).

The progress of the 'Targets 2000' exercise is shown in Table 2.

Table 2 Progress of 'Targets 2000' exercise.

Target area	1999 results	Teachers' initial target (October 1998)	Final target	Progress check (May 1999)	Progress check (January 2000)	Actual result (August 2000)
1+ A*-G grades Zero failure	95%	98.93%	98%	100%	99%	98.9%
5+ A*-C grades High level	55%	56.68%	64%	57%	56%	53%
Average Points score Inclusive	39.2	39.34	41	40.75	39.59	40.47

When the teacher-derived subject level targets were cumulated for the 'crucial' 5+ A*-C grades criterion, the figure of nearly 57% was compared to a potential target of 65% based on Key Stage 3 results and 64% using the newly arrived YELLIS data. However, the use of CAT scores of 102+ as a predictor suggested a target of only 43%. When the CAT 'threshold' was reduced to a score of 97 (following an LEA announcement that pupils with this score or better statistically obtain 5+ A* -C grades at GCSE), a target of 64.5% resulted. Sceptics suggested a 'fix' when the three indicators yielded essentially the same target, and the method used by the Deputy Head (Curriculum) of taking all of the A-band and half of the B-band YELLIS pupils as a criterion for 5+ A*-C grades was thought to be rather crude. Interestingly, the following target-setting process (2001 cohort) went in the opposite direction with targets being set by staff which

were higher than the CAT scores would suggest - yet the YELLIS quartile bands indicate a less able cohort (36% in band A in 2000, 28% in 2001). At the time, the sceptical staff suggested that staff had simply raised their target grades 'to keep the Deputy quiet'. However, the final GCSE results, with a record 59% obtaining 5 or more A*-C grades might also suggest that staff recognised pupil abilities that the statistical indicators did not.

YELLIS received a stay of execution largely as a result of an intervention by the Head of Mathematics. In her memorandum to SMT in November 1998, the Deputy Head (Curriculum) wrote:

> Some colleagues felt that the baseline data did not reflect the skills required to achieve at GCSE in their subjects.... However, [the Head of Mathematics] feels that the statistical methods used by YELLIS/ALIS **do** take this problem into account.... If there is a general feeling that YELLIS is not useful, then perhaps we should consider not paying for it. I do think that feeling has arisen as a response to this year's [1998] value-added analysis rather than reasoned argument, and that it might be more appropriate to discuss the future of YELLIS in the New Year (Deputy Head, 1998c School records)

For her, however, accountability was the real issue. In the same memorandum she directly mentions it three times:

> The issue of target-setting has been the first piece of tangible evidence of increased accountability for all staff.
>
> Colleagues' concerns about increased accountability shouldn't really lead to us underestimating pupils progress or directing our energies to bucking the statistical methods.....Should it?
>
> There are still some colleagues who feel that we shouldn't be doing this level of data analysis.... As a management team, I think we need to work to counteract that, to emphasise that the CCR, Learning Co-ordinators, target-setting etc are as much, if not more, about improving opportunities for pupils as about accountability (ibid).

So despite the 'bottom up' process the Deputy Head (Curriculum) persisted with

a target of 64% 5+ A*-C grades and persuaded the Head to submit this to the LEA. Governors were angered by this - they had not approved the target (Governors' Curriculum Committee 1998, School records). Several thought it too ambitious and that if not achieved it would result in bad publicity for the school. Even though they had received training, there was clearly a lack of understanding amongst governors about the whole process as well as significant disunity. This is illustrated in their final 5+ A*-C grades 'target' of '64% plus or minus 3%' which was finally ratified in January 1999 (Governing Body 1999, School records).

Accountability was also the theme of the Deputy's document sent to all staff in May 1999 (Deputy Head 1999, School records) showing how targets had altered through two progress checks. Targets were not only given at subject- and pupil-level as before, but also at teaching group level. Hence, since all staff had access to the timetable, it was easy to identify individual teachers. As table 2 shows, teachers were generally happy with the zero-failure 1+ A*-C target but the others were gradually eroded to approximately the points that teachers had predicted at the start. Pupils who do not fit the pattern became a topic of conversation. Michael, for example, an A-band YELLIS pupil, was predicted by his teachers to achieve E and F grades. The Deputy's comment 'are we setting that pupil the appropriate level of work?' was countered by comments about Michael's behaviour and his unsupportive parents (Deputy Head 1999; Pupil MG's record file 1999, School records)

The final results were very close to the initial targets proposed by teachers, the

final targets and the second progress check, with the exception of the 5+ A*-C category. The final 5+ A*-C result was 3% down on the second progress check but 11% down on the final target. The whole school average YELLIS standardised residual was minus 0.1. In the 'inquest' that followed there was strong feeling amongst many teachers and governors that the final target had been too high but that a group of able but disaffected pupils, mainly boys, had underachieved.

Heads of Department were also to be held more responsible for the consequences of the way in which pupils were grouped. 'Review the rationale behind pupil grouping' was a key issue from the OFSTED inspection in 1996 (Baxendale 1996: paragraph 8). This rather sweeping statement was not entirely justified within the main body of the report and in view of the significant increase in setting that had occurred in the school since the 1994 'Raising Achievement' report (Raising Achievement Working Party 1994, School records). At Key Stage 4, Inspectors linked ability grouping with the high standards reached in many lessons. Groupings in Science were praised but there was criticism in other subjects, for example Mathematics, about some lack of differentiation 'since ability levels, even in ability sets, are wide' (Baxendale 1996: paragraph 40). Thus ability grouping became a School Development Plan issue (Dalestone School 1997, School records) and Heads of Department now submit their grouping criteria and rationale to the Deputy Head (Curriculum) annually.

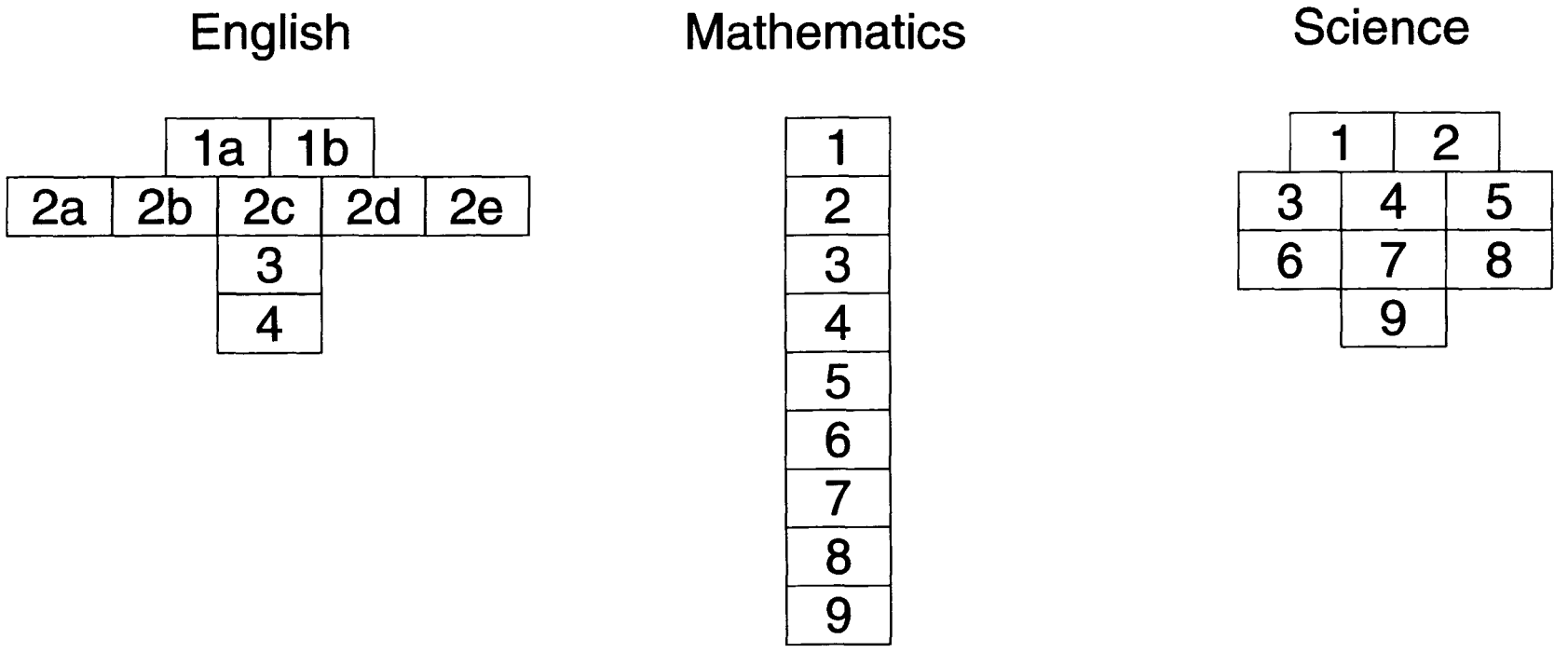
The situation in the core subjects at Key Stage 4 is illustrated in Figure 3 and is

crucial to the interpretation of value-added data to be presented in chapter 5.

The policy in English has changed since 1998 with the appointment of a new Head of Department. Grouping had been based strictly on attainment in English and resulted in top sets dominated by girls. However, the gender imbalance was exacerbated by a policy of placing disruptive boys (irrespective of ability) in lower sets. Now, there are two or three top sets with a better gender balance since aptitude is the most important criterion. The majority of pupils are placed in up to five broad ability 'middle' sets leaving one or two smaller sets providing for those with, for example, literacy problems, some of whom would take the Certificate of Achievement examination.

In Science, Key Stage three results and, to a lesser extent, teacher recommendation are used to produce four ability 'layers'; the number of sets in each layer varies with the profile of ability in the cohort. One or two top sets aim for the highest grades in the 'higher' tier Double Award examination. The next layer is also targeted at the higher level examination whilst that below will aim for the Double Award at 'foundation' level. Some of these students may take the Single Award foundation level examination and this was an option at the start of Key Stage 4 for a group of pupils who also followed a 'half' GCSE course in Art. Pupils in the lowest set usually follow the Single Award foundation level course or the Certificate of Achievement in Science.

Figure 3 Setting Patterns in KS4



Science has always been keen to 'experiment' with pupil grouping in order to raise achievement. For example, the 2001 cohort had single sex top sets from 1999 to 2000 which was principally an attempt to raise the achievement of boys. After one year internal test results showed that few pupils were benefiting and staff had serious reservations about the management of single sex groups in a mixed school. The experiment was abandoned and the sets reorganised for the second GCSE year. The 2002 cohort (with a different ability profile) has been organised to produce an 'express' group which is aiming at A* and A grades.

The Mathematics Department has always created sets in a strictly linear fashion based on attainment. Internal assessments were used until 1995 when confidence in Key Stage 3 results improved. Sets 1 and 2 are directed towards the higher papers, the next three sets follow the 'intermediate' programme, leaving the bottom three taking the foundation level examination. Movement between sets is possible if pupils demonstrate significant improvement or regression, though in practice few changes occur.

By 1999 YELLIS was now firmly back on the school agenda. This was partly driven by one of the Senior Teachers who had attended a YELLIS/ALIS conference and who reported to SMT:

> We have not used YELLIS/ALIS to its full potential. We now have 4 years' of data which should give us ample evidence to identify trends and not to be distracted by natural 'blips'. There are a number of Heads of Department who are either unfamiliar with how to use these to best advantage or need some training, whilst others are keen to do so but have not had access to the full data available. (Key Stage 5 Learning Co-ordinator 1998, School records)

The Deputy Head (Curriculum) then started using the YELLIS bands as a broad

indicator of potential in her accountability 'drive' whilst some Heads of Department pointed to the use of YELLIS as a tool in target-setting. There was new interest in the use of 'chances graphs'. The need to maintain the motivation of pupils had been mentioned in the Value Added National Project Final Report (Fitz-Gibbon 1997: section 4.3) that also asserted the value of such graphs. The Head of History at Dalestone was one of those who were unhappy about the pressures placed on pupils by giving them *specific* targets:

> I did argue at the Heads of department meeting - Autumn 1998 - that we should retain the knowledge so as not to pressurise pupils. A broad range is a more flexible tool. E.g. to tell pupils that we expect them to achieve one of the top three grades is more user friendly and encouraging than to tell a committed pupil "your target is a 'B' ". It demoralises some and has the danger of becoming a self-fulfilling prophecy for others (Dalestone School 1999, School records).

Heads of Department now have a key role in data analysis as was prescribed by Chalkside in 1997, but crucially, value-added features strongly on their agenda as, for example, their commentaries on 1998 and 1999 examination results show:

> It seems clear that we shall have a significant negative residual in the YELLIS analysis. (Geography, 1998: actual residual was -0.4) (Heads of Department 1998, School records)

> High achievement for the YELLIS A/B bands. 6 were YELLIS Band C with a chances graph prediction of C-E; disappointed with the number of D grades. (GCSE PE, 1999) (from KS4 Reports 1999, School records)

Yet perhaps the greatest indicator of the change in the 'climate' is that such reports now frequently identify individual classes - and hence individual teachers. The change comes partly from the 'push' by the Deputy Head (Curriculum) and partly from 'give' on the part of middle managers. There are also clues to a wider acceptance amongst teachers from a survey that was

conducted (see table 3) in May 1999 as part of an audit for Investors In People.

Table 3 Results from Investors In People Staff Survey

18-05-99. 94% response. Figures are percentages.

Question	Agree	Disagree	Not sure
1. I am familiar with the general thrust of the Government's recent proposals with respect to appraisal.	40.9	24.2	34.8
2. I am in favour of a scheme of appraisal that involves classroom observation and other objective evidence of performance.	39.4	37.9	22.7
3. I am in favour of a scheme of appraisal that takes pupil progress into account.	22.7	48.5	28.8
4. I am in favour of being set my own targets at least one of which is directly linked to the school's pupil performance targets.	30.3	43.9	25.8
5. I am in favour of the principle of performance-related pay for teachers.	6.1	86.4	7.6

This shows that around one-third of the teachers agreed with the basic appraisal principles as given in the Green Paper with a further quarter 'wavering'. However, the vast majority at this stage were against the idea of performance-related pay.

4.5 2000-2001: Value-added - the pupil progress measure

The pupil progress standard of the Threshold Assessment applications dominated the practical use of value-added at the school during the first half of 2000. The guidance notes for this standard suggested ways in which teachers could demonstrate the progress of pupils using assessment data in comparison to prior attainment. For the majority of staff YELLIS and ALIS provided the data source of choice not only because it was readily accessible and understandable but also because they had confidence in it.

Out of 44 eligible teachers, 30 applied for Threshold Assessment. Of those not applying, illness and leaving the school were the main reasons offered; only two openly stated that they were simply not applying. Of interest is that a number of those not applying were part-time colleagues. As Deputy Head (Staff Development) I attended a DfEE training day and played a key role in guiding and supporting colleagues through the application process - as well as providing some value-added statistics which had initially been calculated for this thesis. For the pupil progress standard alone, teachers had to provide 2 or 3 specific sets of information giving:

- ◊ a brief context
- ◊ baseline information
- ◊ information about how progress was monitored
- ◊ what progress was expected
- ◊ the actual progress made
- ◊ an evaluation of the process (DfEE 2000:5)

Although the quality of applications varied widely, the most consistent feature was the use of value-added data based on YELLIS or ALIS as the example below illustrates:

> In the last academic year (98-99) I taught 2 Yr 11 groups Chemistry. In each case I was one of three members of staff per group who each taught Biology, Chemistry and Physics.
> 11Sc1 did well. The standardised residual (from YELLIS) for the whole group was +0.55 which is significant. Indeed some pupils were outstanding achieving up to 2 grades better than YELLIS was suggesting.
> 11Sc3 also did well. They were a lower ability set and I did a lot of work on revision and motivation towards the end of the course. The set had a standardised residual of +0.10. However this includes a pupil with long term absence who if he is taken out of the data then the standardised residual becomes +0.16. Whilst this is not a major increase, it is probably a truer reflection of the group's performance.
> (Extract from the pupil progress section of a Science Teacher's Threshold Application, June 2000, with permission. Science Teacher 2000, School records)

The second part of this example illustrates the level of confidence that some teachers now have in using value-added data. It also indicates that guidelines need to be set about what is acceptable in terms of exclusion of data. In a comment about this matter in relation to school-level data, Fitz-Gibbon (1997: section 5.7) said 'the danger is that the pressures of publication will motivate schools towards providing the most favourable view of the data possible'. The pressures of Performance Management might motivate teachers in a similar way.

From many hours of conversations with teachers about Threshold Assessment, I have formed the opinion that this process has done more to encourage reflection on practice than any other initiative, including TVEI and the National Curriculum, that I can recollect in 24 years of teaching.

4.6 Concluding comments

The 'pupil progress' aspect of the Threshold Assessment brought a surge of interest from teachers in value-added data. However, it was clear that, despite the publications of the Deputy Head that had identified individual classes, many teachers had not taken a deep and personal interest in 'their' data. The fact that residuals now had a personal, and potentially financial, significance demanded a 'fresh look', and whilst some teachers were pleased at the revelation that they had 'good' value-added scores others were disappointed. I recollect asking one colleague (also a good friend and with a reputation as a 'good' teacher) whether he would be making a Threshold application. To my surprise he said 'no' as all

of his teaching of GCSE Business Studies classes in recent years had been to lower sets that had obtained negative residuals. Several other teachers, some of whom were subsequently graded as 'very good' or 'excellent' in the 2000 OFSTED inspection found themselves in the same position.

This paradox - where the classes of 'very good' (as graded by OFSTED) teachers obtain negative value-added scores - lies at the heart of this research. To what extent do value-added measures reflect teacher effectiveness?

In the following chapter, the value-added data for core subjects are examined in detail. A unique feature of this analysis is where the residuals for each set are placed side-by-side revealing interesting patterns in pupil performance from top to bottom sets.

Chapter 5 Value-added data

> *...not only does cooking mark the transition from nature to culture, but through it and by means of it, the human state can be defined with all its attributes... (Lévi-Strauss 1970:164)*

5.1 Introduction

In this chapter the Key Stage 4 value-added data for English, Mathematics and Science at Dalestone School from 1995 to 2000 will be analysed. The analysis will start at the subject level and continue at the level of the individual set.

The analysis at set level will explore patterns over time and in relation to the setting policy within each subject area. Since each set is usually the responsibility of one or a small number of teachers, then consideration will be given to the extent to which the analysis will reflect the performance of teachers. A second level of analysis will explore an alternative approach to the interpretation of set-level data. This provides a fairer way of interpreting such data in relation to the performance of teachers and has important implications for Threshold Assessment.

The results from 3,855 GCSE examination results over a period of seven years have been analysed. This represents an average of 184 student results per year. This is smaller than the total number of students on roll (usually between 190 and 200) because some were not present for the YELLIS test or were not entered for GCSE. A small proportion of (usually Special Educational Needs)

children each year are entered for alternative examinations such as Certificate of Achievement. The absentees and non-GCSE entries have been excluded from the data set.

5.2 Subject Level Analysis

The charts shown in figures 4 to 8 show the average standardised residuals for the core subjects over the seven-year period. Each point represents a full cohort, on average 184 GCSE results. Science has been separated into Double Award, Single Award and Triple Award GCSE courses. 'Triple Award' is the designation given to the course followed by one set of more able pupils for three of the years covered by this study. It led to three GCSEs in the separate sciences, that is Biology, Chemistry and Physics. The majority of pupils followed the Double Award course as is now typical in most schools (Donnelly and Jenkins 2001:122).

These are 'statistical process control' (SPC) charts and are used by schools to monitor the progress of each GCSE course with the passage of time. SPC charts superimpose confidence limits on the data to show the variation of results which could occur because there is a new cohort of pupils each year. Confidence limits of two standard deviations give a 95% confidence interval which the CEM Centre describes as a 'warning light', whilst three standard deviations gives 99.7% confidence (Curriculum Evaluation and Management Centre 1999:29).

Figure 4 Average standardised residuals for Mathematics

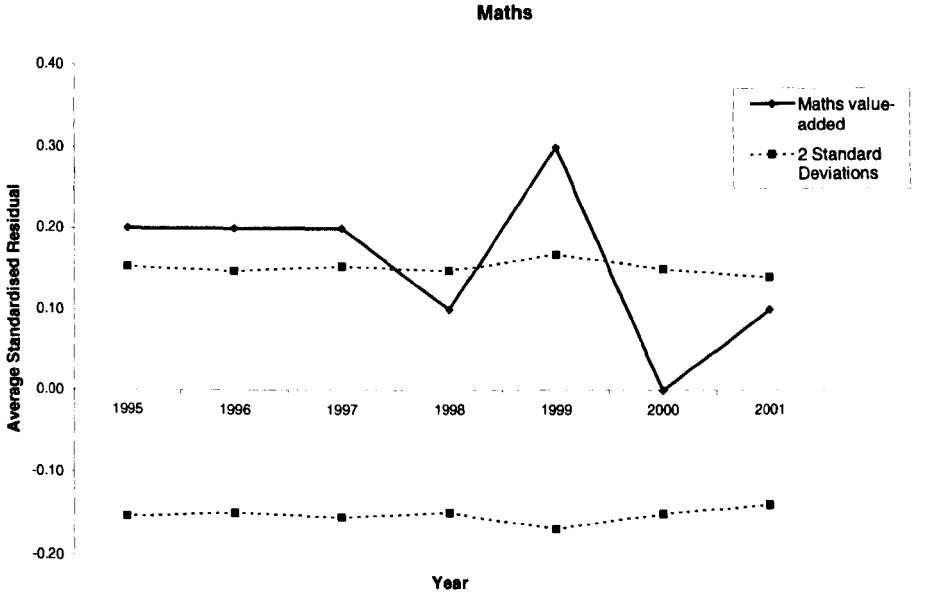


Figure 5 Average standardised residuals for English Language.

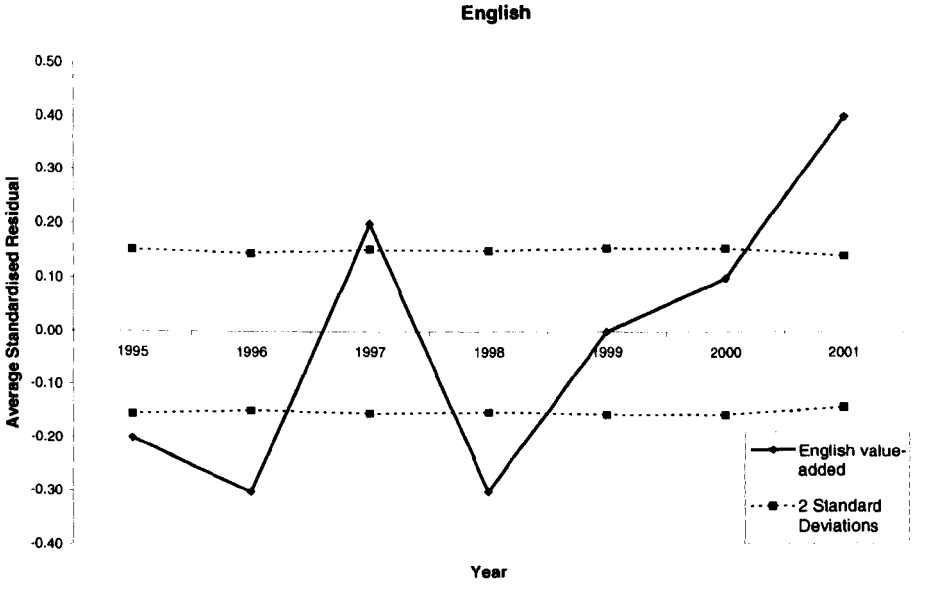


Figure 6 Average standardised residuals for Double Award Science

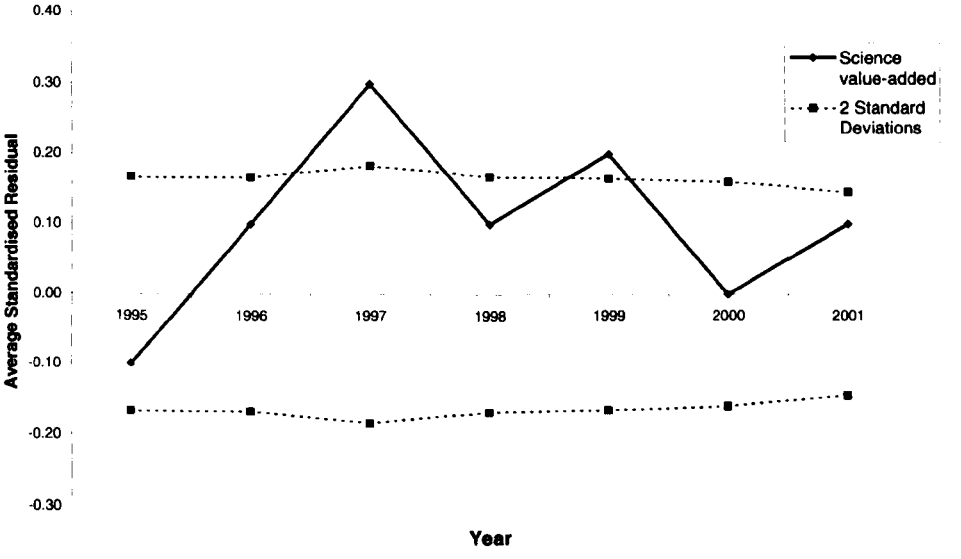


Figure 7 Average standardised residuals for Single Award Science

(There were no single award entries in 1995)

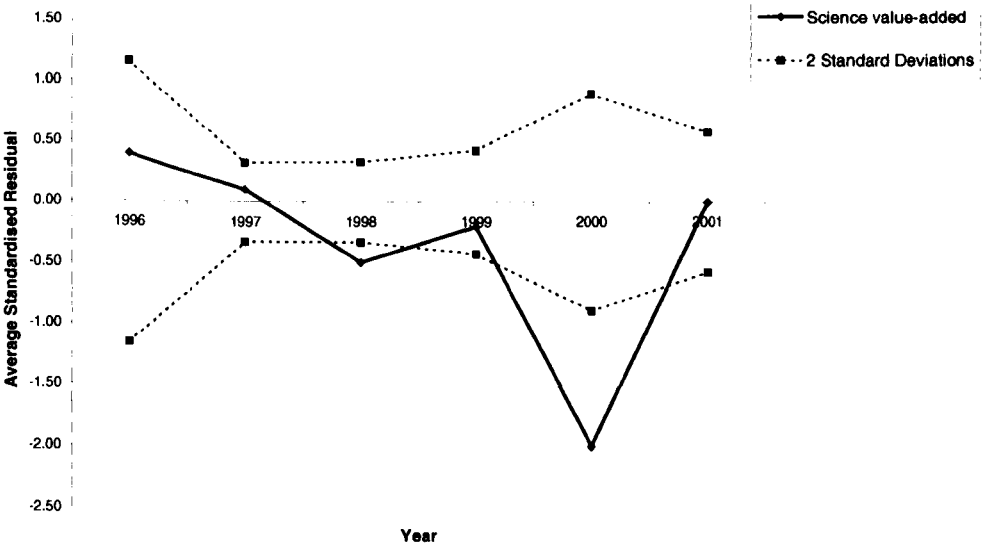
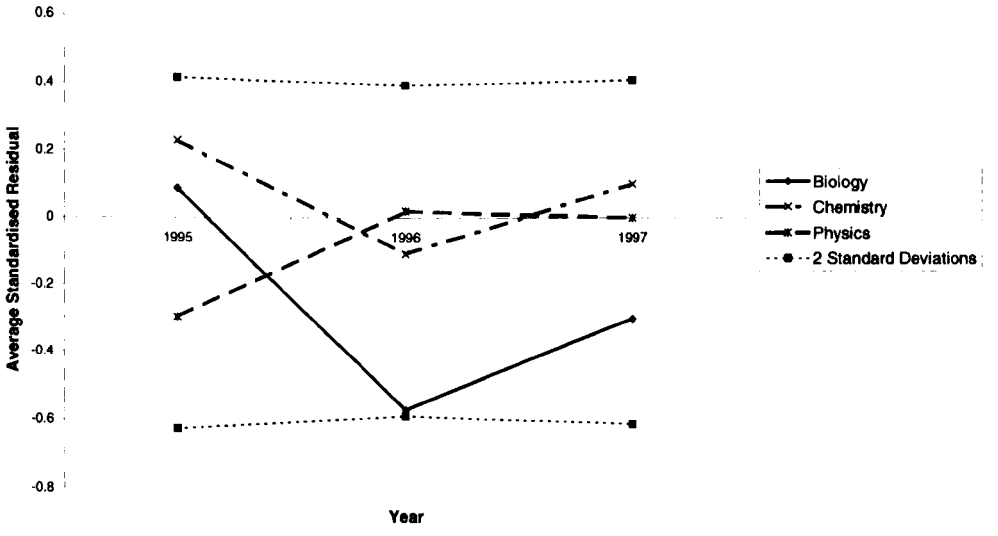


Figure 8 Average standardised residuals for Triple Award Science

(The Triple award was only taught for three of the years covered in this study)



For four out of the seven years the Mathematics results have been significantly better than other most schools in the YELLIS cohort. In the other three years the average standardised residuals are within the control lines indicating performance as might be expected from the baseline YELLIS test.

English language, however, has a more volatile profile with a 'good' year (1997) amongst three 'poor' years where the average standardised residuals were significantly worse than the majority of schools in the YELLIS cohort. The last three years show a notably improving trend which coincides with the appointment of a new Head of department, changed setting arrangements and the introduction of new pedagogical practices.

Double Award Science results are generally within the control lines with two 'good' years. The pattern in recent years is similar to that shown by mathematics. Until 2001, Single Award appeared to be on a downward trend probably reflecting changes in entry policy; although the small numbers of entries must be taken into account - there were only 5 candidates in 2000. The Triple Award chart demonstrates pupil achievement generally in line with similar students in the same subjects in other schools. However, the actual GCSE grades (particularly for Biology) were disappointing for the school, considering that these were the most able students in Science in each of the three years shown. The Triple Award course was not offered to the Year 10 cohort starting in 1996 and so there are no Triple Award results after 1997. This issue will be discussed further in the next section.

The erratic lines on some SPC charts (e.g. English Language) can hide real trends and, more importantly, such charts give no information about the performance of individual sets.

5.3 Analysis of data by sets

5.3.1 Mathematics

Figure 9 shows typical results for Mathematics by set. This is a disaggregation of the +0.2 average standard residual for 1996 and shows that, on average, students in the top three sets obtained significantly positive residuals. Those in the next three sets performed in line with predictions and those in the lowest

sets performed worse than expected, set 7 significantly so. The chart clearly illustrates that an average residual for a subject hides a wide range of results for individual sets.

The data for all years is shown in figure 10 and serves to reinforce the general pattern in figure 9. It reflects the linear way in which pupils are set for Mathematics based on Key Stage 3 National Test results (raw scores) modified by teacher judgement. It is important to note that as a matter of policy there is a rotation of teachers so that one may teach a set 2 one year, followed by a set 7 then a set 4 in the third year.

Figure 9 Standardised residuals for Mathematics sets 1996.

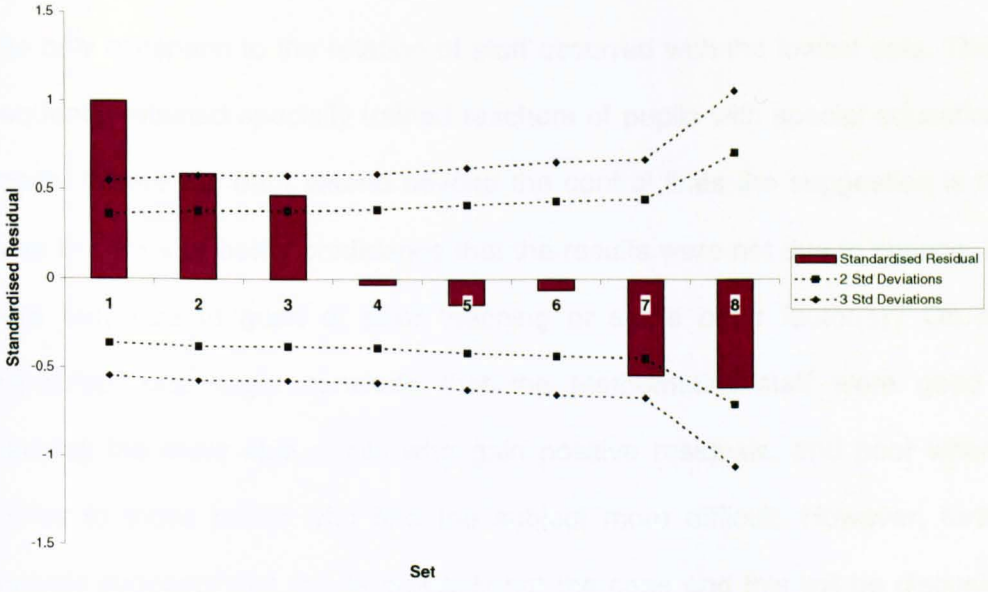
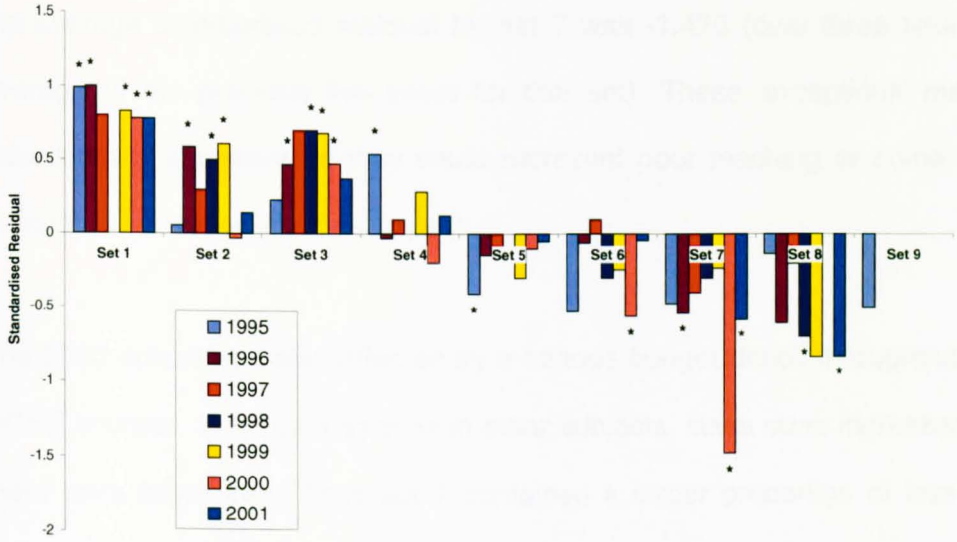


Figure 10 Standardised residuals for Mathematics by set, all years.

Asterisks on this and subsequent charts indicate those sets where average residuals are significant at the 95% level.



The only exception to the rotation of staff occurred with the lowest sets. These frequently retained specially trained teachers of pupils with special educational needs. Where the bars extend beyond the control lines the suggestion is that there is a 95% or better confidence that the results were not due to chance, but were they due to good or poor teaching or some other factor(s)? On first inspection, one might conclude that the Mathematics staff were good at teaching the more able pupils who gain positive residuals, and poor when it comes to those pupils who find the subject more difficult. However, further analysis suggests that this is probably not the case and this will be discussed later.

Although there is a general pattern it is important to note that there are ways in which this is broken. Set 1 in 1998 had an average standardised residual of 0.0; in 2000 the average standardised residual for set 2 was -0.03; and also in 2000 the average standardised residual for set 7 was -1.475 (over three times the average of the previous five years for that set). These 'exceptions' may be educationally important as they could represent poor teaching or some other factor that had an educational impact.

The 2000 cohort was also affected by a serious budget deficit throughout their GCSE courses. In Mathematics as in other subjects, class sizes increased and there were fewer sets. Thus set 7 contained a larger proportion of less able pupils than was usually the case and this may (at least partly) explain the particularly large negative average standardised residual.

5.3.2 English Language

There are many similarities between English Language and Mathematics. Figure 11 (1998) shows a typical pattern that also reflects the setting strategy. In this case two parallel top sets took the most able pupils as determined by Key Stage 3 National Tests, modified by teacher judgement. There followed four parallel 'second' sets and a single bottom set containing the least able pupils. Significantly, the then Head of Department used to place the disaffected boys in this set leaving the top sets dominated by girls.

The data for all years (figure 12) again shows that a pattern is reasonably consistent across time despite a rotation of teachers as in Mathematics, though with specialists again tending to stay with the lowest sets. Top sets usually gain positive average standardised residuals, second sets tend to oscillate around the zero mark, and the lowest sets consistently have significantly negative residuals. Again, the obvious inference is that teachers are successful with top sets and less so with the lowest but it is important to examine this relationship more deeply.

Figure 11 Standardised residuals for English Language sets 1998.

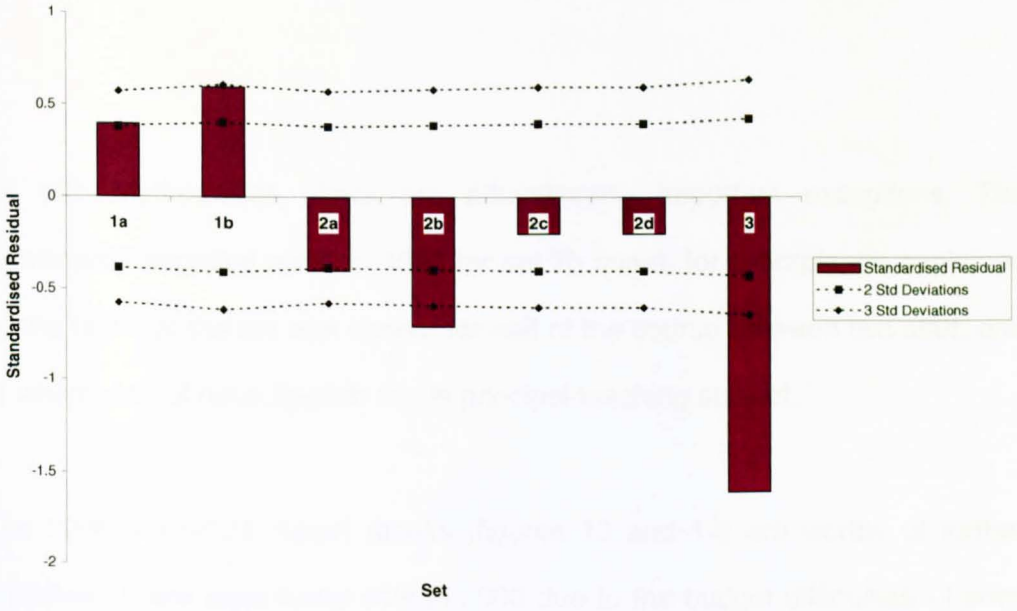
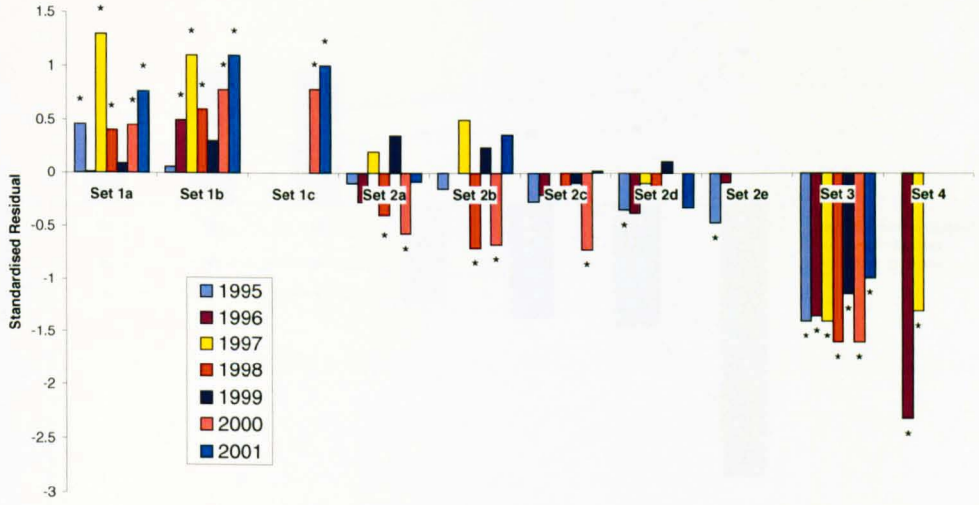


Figure 12 Standardised residuals for English Language by set, all years.



As with Mathematics, there are educationally important exceptions. The significantly negative result in 1998 for set 2b might, for example, be explained by the fact that the set was shared for half of the course between two staff, one of whom did not have English as his principal teaching subject.

The 2000 and 2001 cohort results (figures 13 and 14) are worthy of further attention. There were fewer sets in 2000 due to the budget difficulties - hence class sizes had increased - but of greater significance was the creation of three top sets, three second sets and one bottom set and a better gender balance. This represents the change in setting policy introduced by a new Head of Department that was discussed in the previous chapter.

Figure 13 Standardised residuals for English Language sets 2000.

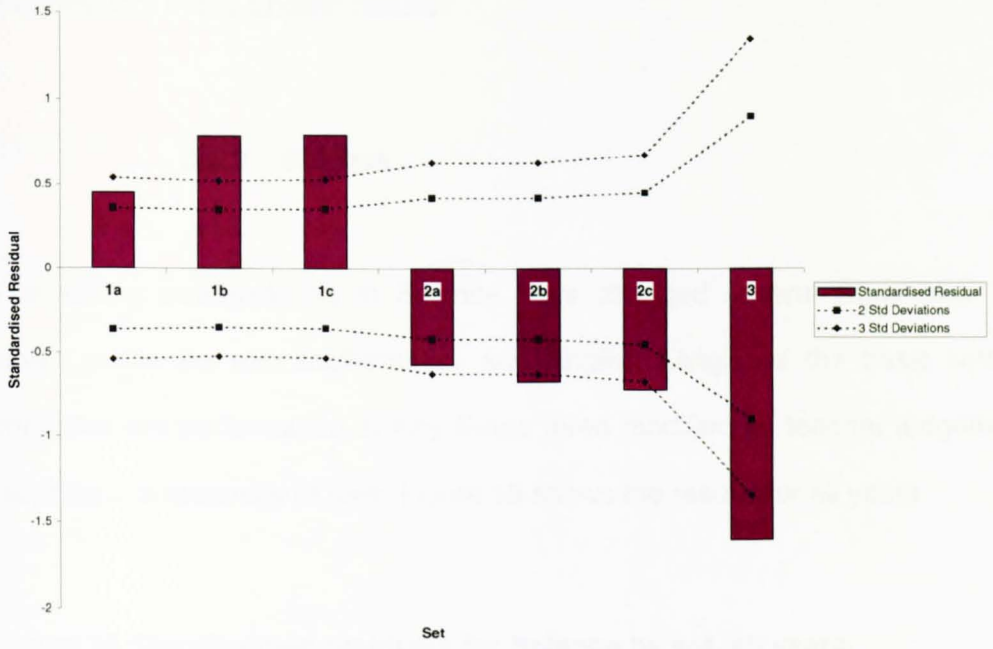
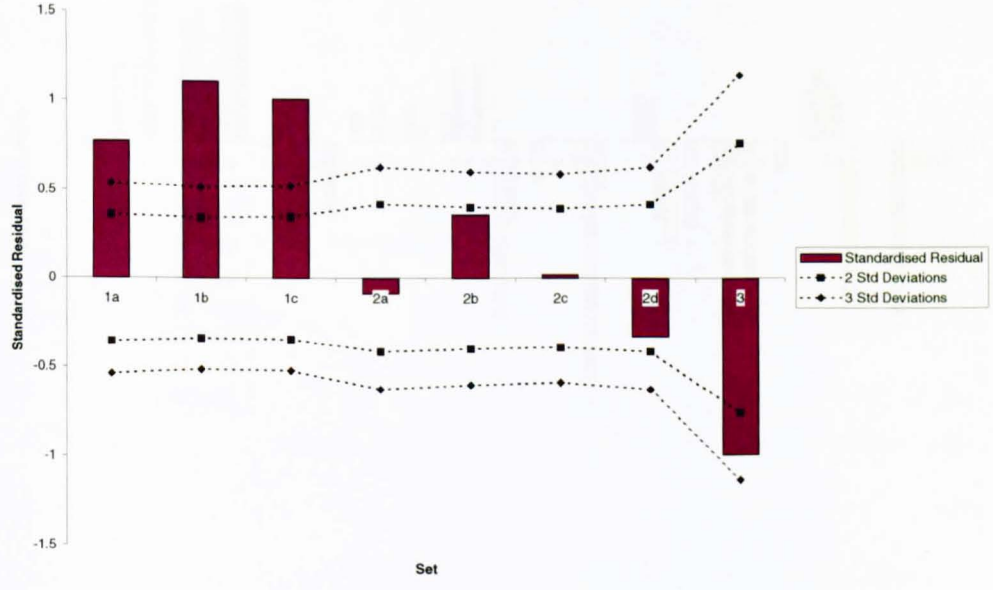


Figure 14 Standardised residuals for English Language sets 2001.



In 2001 there was an extra set and class sizes were reduced. The most noticeable difference though, is the improvement in residuals for the second

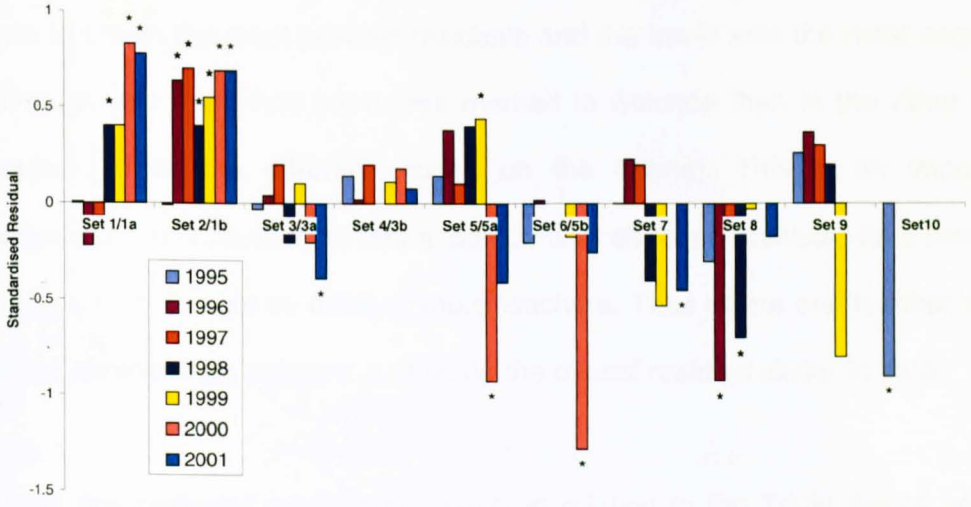
sets, and it is these pupils who have contributed most towards the overall improvement in the English results.

5.3.3 Science

The setting arrangements in Science have changed several times over the seven years. As with Mathematics and English Language the basic setting principles are performance at Key Stage three modified by teacher judgement resulting in a hierarchy of sets. Figure 15 shows the results for all years.

Figure 15 Standardised residuals for Science by set, all years.

(All three Science courses are shown - standardised residuals allow direct comparisons.)



In 1995 pupils were taught in two half-year blocks as a result of a timetabling arrangement. This effectively created two parallel top sets (one in each half

year) and similarly parallel sets 2 to 5. However, one of the top sets was biased towards the most able and this group followed the Triple Award course. This would have implications for the remaining sets in that in terms of ability they would not have been truly parallel.

From 1996 the whole year group was taught at once permitting more flexibility in setting arrangements. The Triple Award course continued in 1996 and 1997 with the remaining sets arranged in a linear fashion as in Mathematics. A smaller SEN bottom set was created. From 1998 the policy has generally been to create pairs of parallel sets but there have been exceptions to this, for example to maintain a discrete SEN set.

Once again (though with important exceptions) there is a tendency for the upper sets to obtain the most positive residuals and the lower sets the most negative although this trend has been less marked in Science than in the other core subjects (note the different scales on the charts). This is an important observation and could represent a 'dampening' effect on residuals as a result of classes being taught by three or more teachers. Thus where one teacher 'adds value' another may 'subtract' it bringing the overall residual closer to zero.

There are particular exceptions notably in relation to the Triple Award groups and set 9 from 1995 to 1998. Thus figure 16 shows the situation in 1996 and figure 17 following the decision to drop the Triple Award course. The reduction in the number of sets in the 2000 cohort has resulted in a pattern that is very similar to that seen in Mathematics and English Language (figure 18).

Figure 16 Standardised residuals for Science sets 1996.

(1B, 1C, 1P = Triple Award Biology, Chemistry and Physics components respectively.)

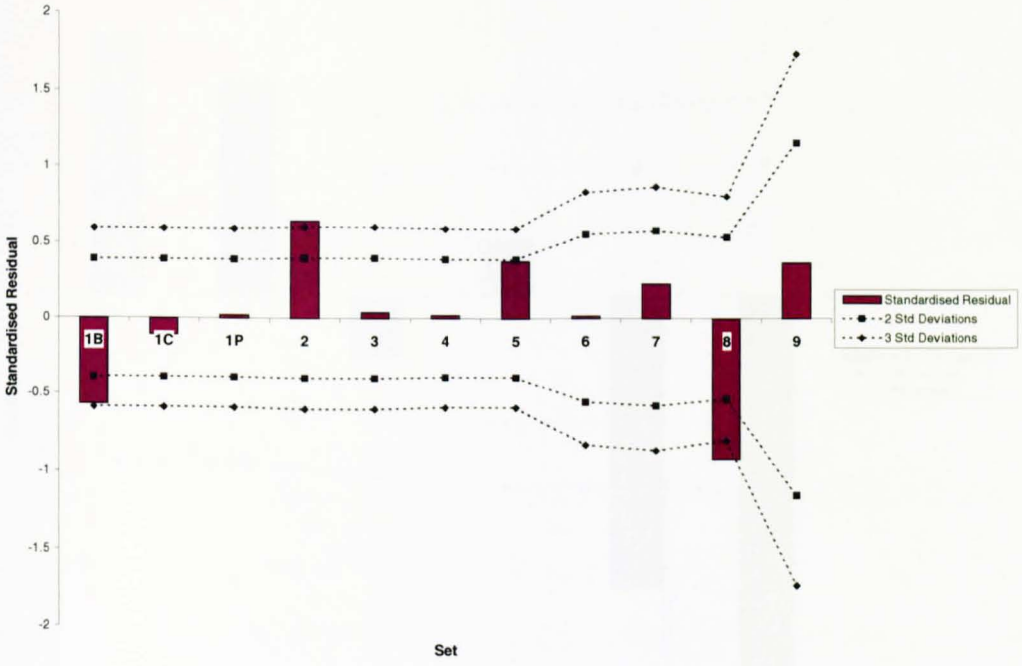


Figure 17 Standardised residuals for Science sets 1998.

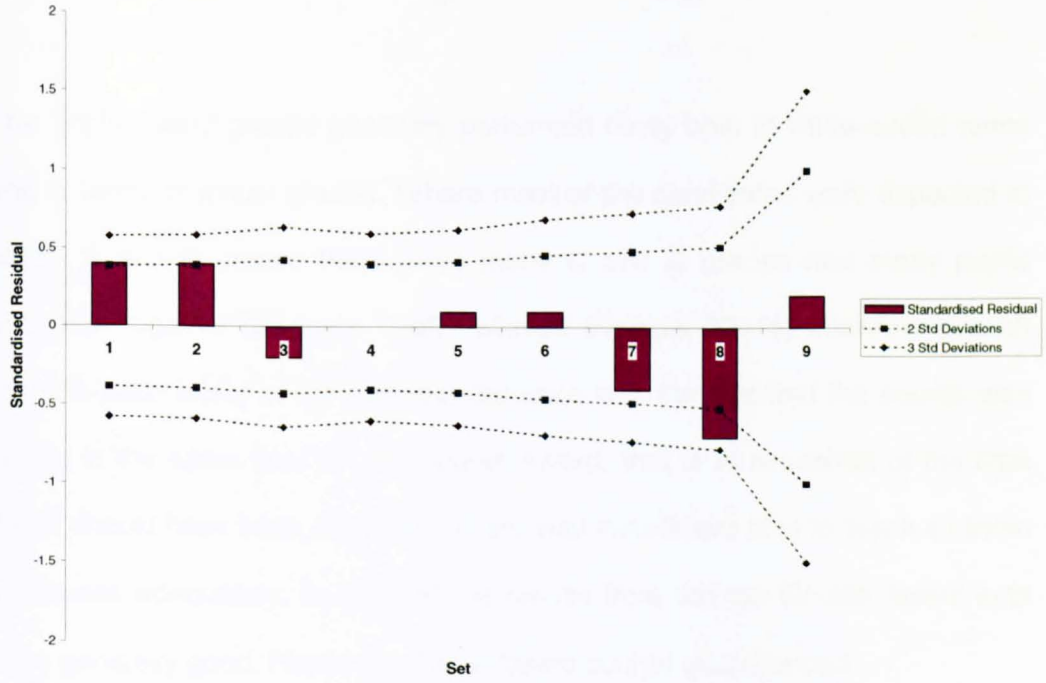
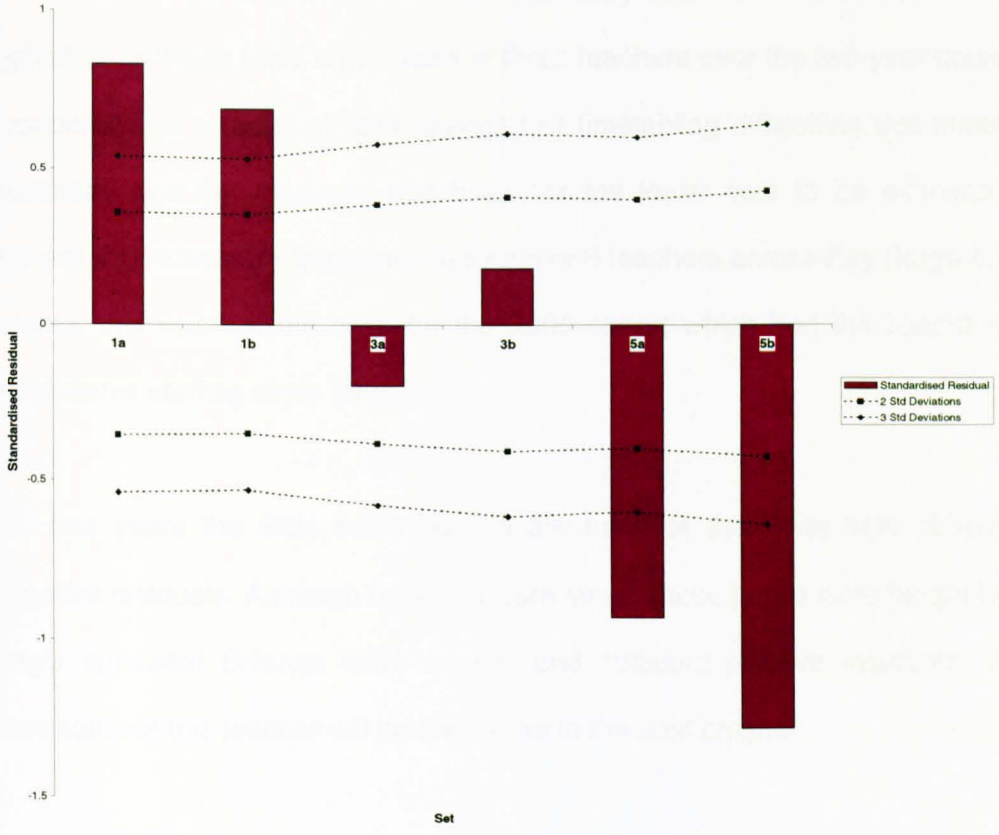


Figure 18 Standardised residuals for Science sets 2000.



The Triple Award groups generally performed badly both in value-added terms and in terms of actual grades. Where most of the candidates were expected to obtain A and B grades there were many C and D grades and many pupils obtained negative residuals. Staff believed (Rogers 2001a, interview record) that the main factor in the poor performance was the fact that the course was taught in the same time as the Double Award, that is in two-thirds of the time that it should have been allocated. There was insufficient time to teach all three syllabuses adequately. In contrast the results from the top Double Award sets were generally good. Hence the Triple Award course was dropped.

Science differs from the other core subjects in that the policy has been to provide classes with specialist Biology, Chemistry and Physics teachers. Thus a typical class would have a minimum of three teachers over the two-year course. Frequently, as a result of staff changes or timetabling difficulties this number increased and the tendency has been for the lower sets to be exposed to highest staff turnover - frequently having 5 or 6 teachers across Key Stage 4. Of interest this was not the case for the 2000 cohort which had the fewest and most stable staffing since 1995.

For four years the SEN set 9 bucked the trend of the lower sets obtaining negative residuals. Although numbers were small, these pupils were taught by a single specialist Science SEN teacher and obtained positive residuals. The comments of this teacher will be discussed in the next chapter.

Thus in Science there are clues that teachers may make an impact on the residuals recorded for the sets that they teach. The instability in Science staffing has given many more average standardised residuals for sets that are close to zero as illustrated in figures 16 and 17, than tend to be found in Mathematics and English. Further, a SEN set with a single teacher can obtain positive residuals when the general trend is otherwise. Finally, when there is more stable staffing and when the pupils are squeezed into fewer sets then a pattern of value-added scores has developed that is in line with other subjects.

5.3.4 Issues arising from the analysis of data by sets

In considering the data, a number of variables have been identified which are likely to have an effect on the residuals. These include strategic factors such as setting policy and the impact of the school budget (for example on class size). The deployment of teachers in terms of whether sets are shared or not, staff changes, the timetable and the pupil composition of sets (by gender or behaviour) are amongst other variables. The extent to which each has an impact on residuals is beyond the scope of this study. It is to be expected that there will be a complex of factors operating in different, perhaps contradictory ways in each set.

Significantly though, a pattern of top sets obtaining positive residuals, modest residuals in the middle sets and negative scores for the lowest has been recognised in all of the core subjects. It has also been found in other schools. Figure 19 shows sample results for a school in the West Midlands (School A 2000, School records) and figure 20 shows sample results for a North Yorkshire school (School B, 2000, School records).

Figure 19 Standardised residuals for School A (Science, 1998)

Sets were designated High (H), Medium (M) or Low (L) ability or a combination of these.

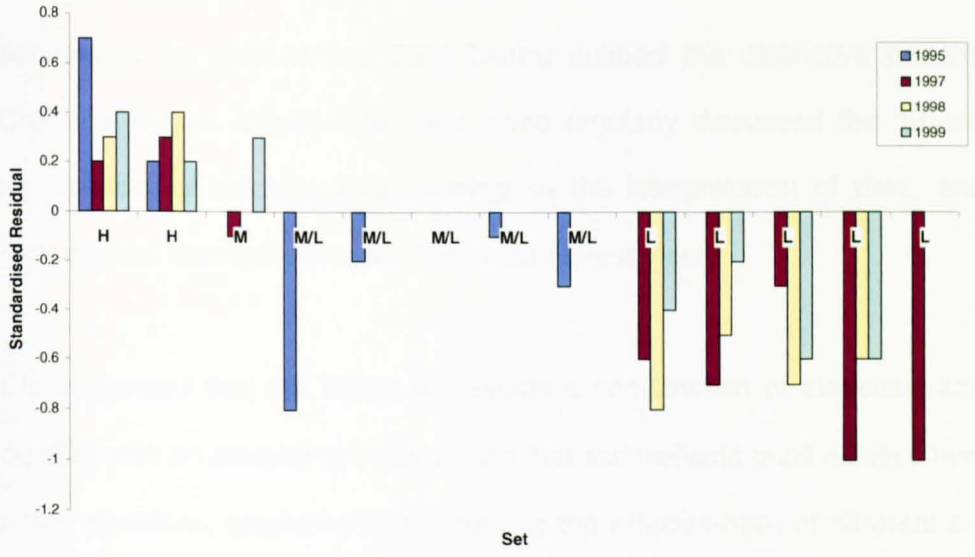
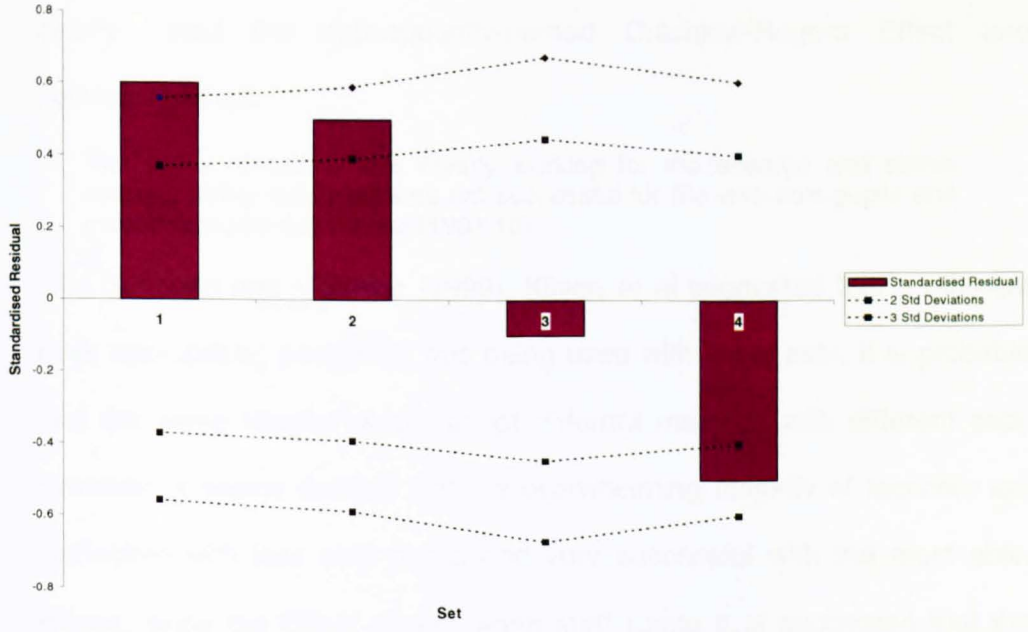


Figure 20 Standardised residuals for School B (Mathematics, 1999)



The pattern was discussed with staff at the CEM Centre in Durham from July 2000 (Rogers 2001b, interview record). It had simultaneously been recognised by John Critchlow, a North Yorkshire Headteacher. As a result of these discussions the staff at the CEM Centre dubbed this distinctive pattern the 'Critchlow-Rogers Effect'. We have since regularly discussed the 'Effect', its significance for teaching and learning, in the interpretation of data, and for Performance Management and Threshold Assessment.

It is suggested that the Effect represents a combination of statistical artefact together with an amount of value-added that *truly* reflects pupil achievements in setted situations, teacher effectiveness or the effectiveness of different setting policies. Seven features of the Critchlow-Rogers Effect led us to believe that it is substantially a statistical artefact:

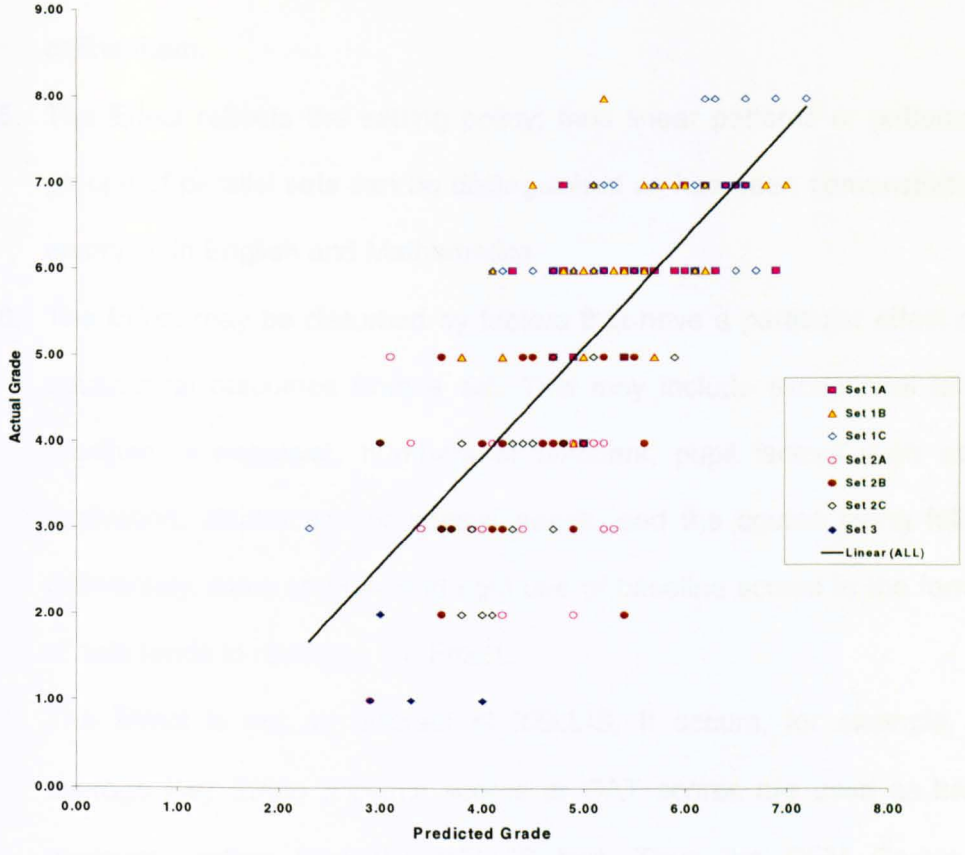
1. The Effect occurs despite rotations of teaching staff. In 1997 Kilyon *et al* had clearly noted the subsequently-named Critchlow-Rogers Effect and commented that:

   > The policy of setting was clearly working for the average and above average ability pupils but was not successful for the less able pupils and indeed could be detrimental (1997:10).

   Like Schagen and Morrison (1999), Kilyon *et al* suggested that a different (less appropriate) pedagogy was being used with lower sets. It is probable that the same teacher *would* adopt different methods with different sets. However, it seems unlikely that the overwhelming majority of teachers are ineffective with less able pupils and very successful with the most able. Hence, since the Effect occurs when staff rotate it is suggested that the reason for the Effect is at the very least not entirely due to pedagogy.

2. The same Effect occurs in different subjects which suggests that it is not specifically related to a feature of one subject, its syllabus requirements or how it is taught.
3. The Effect occurs in different schools - it is not therefore unique to Dalestone and related to one school's organisation.
4. The Effect occurs as an outcome of setting pupils. Data would need to be obtained from schools where pupils were taught in mixed ability sets in core subjects to confirm this. However, when data for a given cohort are plotted in the form of a scattergram the result, a scatter of points above and below the line, is typically as shown in figure 21.

Figure 21 Scattergram showing predicted grades vs actual grades for English Language (2000).



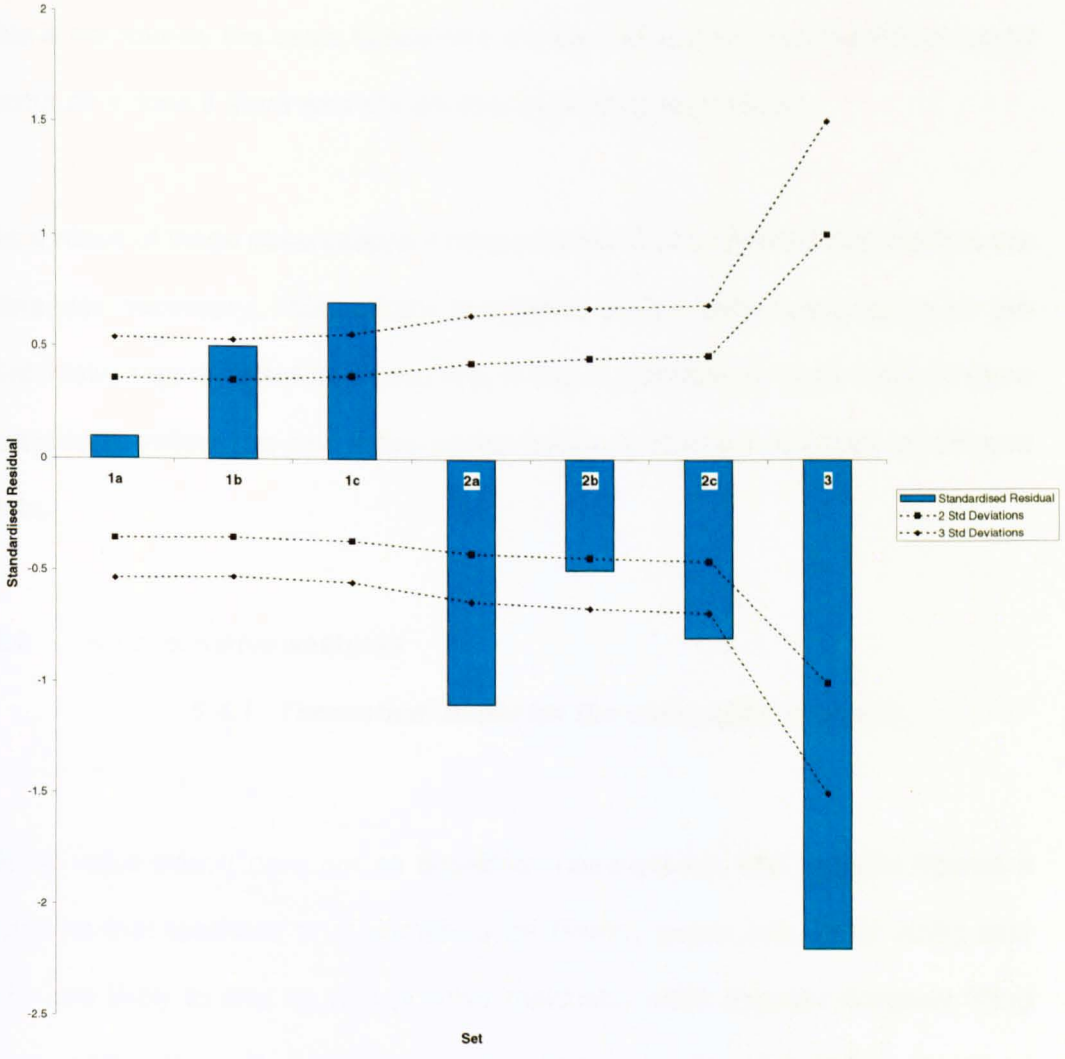
At Dalestone School the only mixed ability sets occur in, for example, Humanities subjects at Key Stage 4. Since these subjects are optional they do not attract a whole cohort and there may be some bias in these 'mixed ability' sets due to patterns of option choices. Hence no appropriate mixed ability data from this school is available.

Where setting occurs, students in the top set tend to be represented by points in the top horizontal 'slice' where most points lie close to or above the line. Thus, on average a top set has a positive average value-added score. For the bottom set the opposite is true. If students were placed in 'baseline' sets, average residuals would generally be close to zero as vertical 'slices' define them.

5. The Effect reflects the setting policy; thus linear patterns or patterns with groups of parallel sets can be distinguished as has been demonstrated, for example, in English and Mathematics.
6. The Effect may be disturbed by factors that have a particular effect on the educational outcomes from a set. This may include exceptional teaching (positive or negative), numbers of teachers, pupil factors such as high motivation, disaffection or special needs, and the course being followed. Conversely, more stability and rigid use of baseline scores in the formation of sets tends to reinforce the Effect.
7. The Effect is not an artefact of YELLIS. It occurs, for example, when average Key Stage 3 points scores or CAT scores are used as baseline measures rather than the YELLIS test. Thus the CEM Centre have

calculated new standardised residuals for the same cohort shown in figure 13 (English Language 2000) using average Key Stage 3 points scores as the baseline. The resulting value-added chart, shown in figure 22, is remarkably similar to figure 13.

Figure 22 English Language 2000 cohort. Residuals calculated using KS3 points as the baseline.



The key issue is that a superficial view of the residuals obtained in setted situations strongly suggests that top sets 'overachieve' whereas lower sets 'underachieve'. A simple explanation for this would be to say that this is due to the teachers and the methods that they employ. It might be argued that top sets receive better resources or the pupils are more highly motivated. Any, some or all of these are logical explanations. However, the combination of the seven observations above offers, in a legalistic sense, a 'balance of probability' that this is not entirely the case. In essence it is hard to imagine that the Effect would occur as it does if there was not another overriding explanation.

As a result of these observations it became clear that a different approach to the data was necessary. This is firstly in order to seek a better explanation for the Critchlow-Rogers Effect and secondly, if this is possible, to seek ways to more realistically reflect the *true* value-added measure attained by pupils in different sets.

5.4 An alternative analysis

5.4.1 Theoretical basis for the alternative analysis

Since value-added data are all based on comparisons with national figures it appears that teachers must somehow be filtering pupils into upper ability sets who are likely to end up with positive residuals, while disproportionately filling lower sets with pupils whose residuals will turn out to be negative. Heads of Department would deny this purely on the grounds that they have no idea what future residuals are going to be. However, when setting targets for Year 10

pupils subject staff are making increasing use of a Central Curriculum Record (CCR). This is a database containing all of the important numerical data about each pupil's achievements, for example Key Stage 2 and 3 results, reading and spelling quotients, and the results of Cognitive Ability Tests. The YELLIS test results are not available until the middle of the Autumn Term of Year 10 and so are not available in time for target-setting. The CCR also contains average Key Stage 3 test scores which Heads of Department tend to make little use of - they prefer to use the results for their own subject since the average Key Stage 3 score might mask strengths or weaknesses in that subject. This is perhaps the most crucial part of the Critchlow-Rogers Effect - the use of a broad *baseline assessment in the calculation of value-added* but the use of *subject-specific data in the creation of sets.* It must be emphasised that this is only a partial explanation as there is strong evidence from teacher interviews to suggest that other factors combine with and exaggerate or dampen this statistical effect. These will be discussed in subsequent chapters.

In order to explain how the effect occurs, CCR data for a selection of 9 pupils is presented in table 4. In this and subsequent tables the following colour coding 'zones' have been used:

YELLOW	Pupils who have a reasonably balanced profile of ability across the core subjects.
TURQUOISE	Pupils who are stronger in the subject in question compared to the other core subjects
RED	Pupils who are weaker in the subject in question compared to the other core subjects

Table 4 Sample Central Curriculum Record - early Year 10

Grades are given as letters and points; A*=8 points, A=7....G=1

Name	Key Stage 3 Results			Maths Set	YELLIS Score	Predicted Maths GCSE grade
	Maths	English	Science			
Janine	7	6	6	1	60	B/6
Tom	6	6	6	1	60	B/6
Helen	5	6	6	2	60	B/6
Grant	6	5	5	1	45	D/4
Kate	5	5	6	2	45	D/4
Paul	4	6	5	3	45	D/4
Kaye	5	3	4	2	30	F/2
Adam	4	4	4	3	30	F/2
Lucy	3	5	4	3	30	F/2

It is common experience that many pupils are stronger in some subjects than in others and this is substantially why setting has come to be preferred to streaming or banding. A subject department such as Mathematics will set pupils according to their performance in that subject, typically using Key Stage 3 or internal test results, modified in some cases by the judgements of their teachers. On the other hand, the more broadly based measures of performance are chosen to act as the baseline for value-added calculations, for example the YELLIS test or average Key Stage 3 points scores. So, YELLIS scores and GCSE are not cogeneric, that is, they do not measure the same thing. Hence, set 1 will contain pupils who perform best in Mathematics but some of them will have performed relatively less well in the overall baseline test (such as Grant - shaded turquoise), as this reflects ability in areas where they may not be so strong. Grant is likely to do better in Mathematics than would have been predicted from his baseline score. Similarly, Kaye is placed in set 2 for Mathematics as it is a strength for her; whilst Helen (shaded red) is in set 2 as Mathematics is her weakest core subject.

The bottom Mathematics set will contain some pupils whose baseline scores are rather higher - such as Paul. He will probably do less well in Mathematics than would have been predicted from his overall ability, making a negative average residual a strong possibility when his GCSE grade in Mathematics comes to be compared with the prediction from his baseline score. Other pupils (shaded yellow) generally have a more balanced profile of ability.

In table 5 the pupils have been sorted into their sets and final grades entered.

Table 5 Sample Central Curriculum Record - after GCSE results published.

Grades are given as letters and points; A*=8 points, A=7....G=1

Name	Maths Set	YELLIS Score	Predicted Maths GCSE grade	Actual Maths GCSE grade	Residual	Average residual for each set
Janine	1	60	B/6	A/7	+1	+0.67
Tom	1	60	B/6	B/6	0	
Grant	1	45	D/4	C/5	+1	
Helen	2	60	B/6	C/5	-1	0.0
Kate	2	45	D/4	D/4	0	
Kaye	2	30	F/2	E/3	+1	
Paul	3	45	D/4	E/3	-1	-0.67
Adam	3	30	F/2	F/2	0	
Lucy	3	30	F/2	G/1	-1	

Those who are relatively more able in Mathematics have been presumed to outperform their prediction by one grade, those who are relatively weak achieve one grade lower than predicted, and the rest achieve their predicted grade. The average residuals have been calculated and clearly show the Critchlow-Rogers Effect in that average residuals vary from positive through zero to negative on moving down the ability sets. The top set achieves an average residual of +2/3 of a grade, thus outperforming their potential, the middle set performs as

expected and the bottom set performs below potential by achieving an average residual of -2/3 of a grade. This effect may well mask any variations attributable to the relative progress that the group has actually made. However, it is clearly unfair to use these measurements as they stand to make judgements about teaching competence.

If these pupils had been set according to their baseline YELLIS test score then the average residuals would have been as in table 6.

Table 6 Sample Central Curriculum Record - after GCSE results published but pupils sorted by notional baseline set.

Grades are given as letters and points; A*=8 points, A=7....G=1

Name	Maths Set	YELLIS Score	Baseline set (based on YELLIS score	Predicted Maths GCSE grade	Actual Maths GCSE grade	Residual	Average residual for each baseline set
Janine	1	60	1	B/6	A/7	+1	0.0
Tom	1	60	1	B/6	B/6	0	
Helen	2	60	1	B/6	C/5	-1	
Grant	1	45	2	D/4	C/5	+1	0.0
Kate	2	45	2	D/4	D/4	0	
Paul	3	45	2	D/4	E/3	-1	
Kaye	2	30	3	F/2	E/3	+1	0.0
Adam	3	30	3	F/2	F/2	0	
Lucy	3	30	3	F/2	G/1	-1	

Thus if setting had been based on baseline scores we would expect the average residuals to be zero.

5.4.2 Applying the alternative analysis to data from Dalestone School

This argument will now be explored in more detail through a reworking of the data for a selection of the sets mentioned in section 5.3. In order to aid the discussion colour has been applied to the charts as in the last section. Thus reference to the 'red zone' considers pupils who are weak in a particular subject compared to their baseline scores - these pupils are more likely to obtain negative residuals. The 'turquoise zone' pupils have a strength in the subject and might be expected to gain positive residuals. 'Yellow zone' pupils are the majority who tend to have a more balanced profile of ability and obtain grades closer to predictions, that is residuals close to zero.

In table 7 the data is that which was presented in figure 9. The average standardised residuals across the bottom of the table are those used to plot figure 9. The cohort has been ranked according to their YELLIS baseline score and new 'baseline' sets have been formed, each containing the same number of pupils as were in the actual sets.

Table 7 Mathematics 1996 - analysis of residuals by baseline sets and actual sets

Base -line sets	Actual sets																Ave std Residual (Base-line set)
	1	*1 No*	2	*2 No*	3	*3 No*	4	*4 No*	5	*5 No*	6	*6 No*	7	*7 No*	8	*8 No*	
1	0.9	*19*	-0.1	*10*	-1.2	*1*											0.50
2	1.1	8	0.5	*9*	-0.1	*6*	-0.9	*3*	-0.8	*1*							0.33
3	1.3	*3*	1.1	*3*	0.4	*8*	-0.0	*9*	-0.3	*1*	-0.4	*3*					0.34
4			1.7	*1*	0.6	*6*	-0.1	*8*	-0.3	*6*	-0.3	*3*	-2.8	*2*			-0.14
5			2.0	*3*	0.8	*3*	0.8	*5*	-0.6	*6*	-0.1	*5*	-1.2	*1*			0.30
6			1.5	*1*	1.6	*2*	-1.8	*1*	0.3	*7*	0.2	*5*	-1.2	*5*			-0.01
7					2.1	*1*			5	*2*	-0.1	*4*	-0.0	*9*	-1.2	*4*	-0.13
8											1.3	*1*	0.8	*3*	0.0	*4*	0.46
Ave SR	1.0		0.6		0.5		-0.0		-0.1		-0.1		-0.5		-0.6		***0.21***
No in set		*30*		*27*		*27*		*26*		*23*		*21*		*20*		*8*	

Thus of the 27 in 'actual' set 3, only 8 would have been in baseline set 3. The 7 in the red zone above would have been in higher baseline sets and 12 in the turquoise zone would have been in lower baseline sets. In line with the theoretical analysis all of the residuals in the red zone are negative and all but two in the turquoise zone are positive. Of interest is the -1.8 residual in the turquoise zone for actual set 4. This represents a boy who arrived at the school in Year 10 and experienced great difficulties settling into his new school - another variable to be considered in analyses of this kind.

The average standardised residuals for the baseline sets do not show the pattern that is found with actual sets. Five out of the eight baseline sets are positive, reflecting the overall positive average residual for the subject (0.21), that is, the pupils have on average performed around one-fifth of a grade better than similar pupils nationally.

The yellow zone pupils are in an 'actual' set that is at the same level as their 'baseline' set. Such pupils might be expected to obtain residuals closer to zero if their progress mirrored the national picture, and the majority do. It is suggested that only at this level of analysis might the value-added data be used to convey something about the performance of the teacher. Only in set 1 is the average standardised residual in the yellow zone significant which may suggest an above average performance by the teacher. However, such statements at this level of analysis have to be made with great caution since the numbers of pupils involved are much smaller in this analysis.

This analysis serves to further illustrate the Critchlow-Rogers Effect, particularly in that it is pupils in the red and turquoise zones who contribute most to it. For set 1 there are no red zone pupils, hence a negative residual for this set would have to indicate that there was another factor at work, for example poor teaching or (as was probably the case for the Triple Award Science) lack of teaching time. Similarly, the bottom set has no turquoise zone pupils and again would only get a positive residual under exceptional circumstances.

Table 8 shows the same analysis as applied to the 2000 cohort's English Language results.

Table 8 English Language 2000 - analysis of residuals by baseline sets and actual sets (compare with figure 13)

Base-line sets	Actual sets														Ave std Residual (Base-line set)
	1a	*1a No*	1b	*1b No*	1c	*1c No*	2a	*2a No*	2b	*2b No*	2c	*2c No*	3	*3 No*	
1	0.2	*24*	0.7	*27*	0.8	*25*	-1.5	*7*	-1.7	*8*	-0.7	*5*			0.16
2	1.3	*7*	1.2	*6*	0.9	*6*	-0.0	*15*	-0.2	*14*	-0.9	*13*	-1.6	*4*	0.00
3					0.7	*1*	-1.8	*1*			0.3	*2*	-1.8	*1*	-0.48
Ave SR	0.45		0.78		0.79		-0.6		-0.7		-0.7		-1.6		***0.1***
No in set		*31*		*33*		*32*		*23*		*22*		*20*		*5*	

The pattern of results in the three zones is again in line with the theoretical analysis. As there are broadly two groups of three parallel sets comparisons are easier to make. Thus looking at the yellow zone of sets 1a, b, and c it is clear that the latter two have performed distinctly better than set 1a. Similarly, the yellow zone of set 2c has performed much worse than 2a or 2b. This was not apparent from the average standardised residuals for the 'actual' second sets. The analysis has therefore removed the gross effects of pupils who naturally perform better or worse in English than in the more broadly based baseline test. It has then brought us closer to those pupils who have 'truly' added (or lost) value.

'Truly' is perhaps not quite true. The three zones, like the sets on which they are based, have arbitrary cut-off points and do not represent the continuum of pupil ability that exists. Hence, for example, some pupils in the yellow zone will be 'closer' to the red zone than others and this analysis does not take that into account.

The 2000 Science analysis in table 9 is also broadly in line with the theoretical model but since each set would have been taught by at least three teachers then apportionment of responsibility for positive or negative residuals is not possible.

Table 9 Science 2000 - analysis of residuals by baseline sets and actual sets (compare with figure 18)

Base -line sets	Actual sets												Ave std Residual (Base-line set)
	1a	*1a No*	1b	*1b No*	3a	*3a No*	3b	*3b No*	5a	*5a No*	5b	*5b No*	
1	0.7	*23*	0.6	*27*	-0.7	*8*	-0.8	*5*					0.78
2	1.4	9	1.3	*5*	0.1	*14*	0.14	*14*	-0.9	*7*	-2.1	*3*	-0.02
3					-0.1	*5*	0.9	*6*	-0.9	*18*	-1.2	*18*	-1.09
Ave SR	0.83		0.69		-0.2		0.18		-0.9		-1.3		***-0.06***
No in set		*32*		*32*		*27*		*25*		*25*		*21*	

The analysis of English Language and Science data demonstrates that the bias that was evident in the actual set residuals is not entirely removed in this analysis. The average standardised residuals for the baseline sets are positive for the top baseline set, zero (or close to zero) for the second set and negative for baseline set 3. In the above English example, the average residual of all the pupils who would notionally have been placed in baseline set 2 is zero, showing that their progress mirrored that of similar pupils nationally. Some of them are in actual English sets 1a, 1b and 1c, as they are better at English, and they would therefore be expected to achieve more positive residuals, while those in English sets 2a, 2b and 2c would end up with residuals which are more negative. What this means is that there is still a potential for bias towards positive average

residuals in upper sets and negative scores for lower sets. Where most of the pupils in a notional baseline set have ended up in the same actual set, the effect will be quite small and can be ignored.

In theory then, the bias is thus likely to be greatest in top and bottom groups where setting is linear and in situations where the cohort is effectively split down the middle into parallel groups of upper and lower ability. In the Mathematics example (table 7) this pattern can only be partially discerned and this further illustrates the complexity that exists in education and the many variables that have to be considered.

5.5 Concluding comments

The patterns of residuals for sets generally mirror the policy that is in place for the setting of pupils irrespective of subject, with 'top' sets mainly obtaining positive residuals and 'bottom' sets negative scores. This, the Critchlow-Rogers Effect, has been the most significant finding from the analysis of value-added data. The effect is observed in all of the core subjects and is not restricted to Dalestone School. Further, a set of observations about the Effect strongly suggest that the effect is to some extent (but not exclusively) independent of the teacher. This will be discussed in Chapter 7.

Using a contrasting analysis of the data it seems possible to partially get behind the Effect and therefore closer to the non-statistical factors that influence pupil achievement. Where there are exceptions to the Critchlow-Rogers Effect it is

often possible to suggest a reason for the residual(s) in question. Variables include the effects of setting policy, pupil behaviour, whether sets are shared between staff or not, and the timetable. However, the most important factor is expected to be the teacher. The next chapter will consider the issues arising from teacher interviews and that following will examine the nature of the interaction between value-added and the data of teachers.

Chapter 6 Teachers' voices

> *...the quality of pupils' science education is conditioned above all by the quality of the teaching which they experience, and...the single most important influence on this, by some way, is the teacher (Donnelly and Jenkins 2001:167).*
>
> *...it is difficult to imagine that teachers are not the most important factor in a good education. Unfortunately...the study we have conducted...does not support this view....If differences between teachers are not usually a statistically significant factor then we must rethink many school-improvement issues (Kelley 1999).*

6.1 Introduction - the professional context

Using YELLIS and ALIS data, Kelley found, to his surprise, that there were no significant differences in value-added terms in most of the subject departments in his Tyne and Wear school. He found that the use of revision materials and ICT appeared to make a bigger impact. This contrasts with much published literature and government policy which promotes the notion that 'teachers make a difference'. In this chapter interview data will be analysed to establish teacher views on teacher effectiveness and the extent to which teachers perceive themselves to be 'effective'. In particular their comments on their attitudes towards different sets will be evaluated. The chapter concludes with a consideration of teacher's views on performance-related pay.

Donnelly and Jenkins (2001:120) noted that the views of science teachers 'did not have a major influence in the process of [science] policy creation and, in the field of implementation, their role was seen by government and influential groups...as essentially responsive'. This is also probably true for other

curriculum areas since minimal consultation and strong central prescription was certainly the approach utilised particularly in the early years of the National Curriculum (Helsby and McCulloch 1997:2). Effectively, 'teachers' voices have frequently been silenced by policy' and policymakers have repeatedly ignored or excluded the voices of teachers in the reform process (Hargreaves 1996:12). However, many examples exist from the context of practice, for example the National Numeracy Strategy, to show that despite this

> teachers' own experiences, values and purposes will all play a part as they process what they see, hear and are offered, and make sense of it in their own ways (Brown, Millett et al. 2000:469).

As Bowe and Ball (1992:22) have concluded, 'the key point is that policy is not simply received and implemented within this arena rather it is subject to interpretation and then 'recreated'.

Thus this analysis considers teachers' self-image, what it is to be a professional and what it is to be 'effective'. These notions will inform teachers' views of value-added as an initiative this will be explored in greater depth in the next chapter.

The majority of the teachers who were interviewed have had professional lives that appear to have evolved gradually. In some cases parents or relatives were teachers and it seemed natural that they would follow in their footsteps. In other cases a positive experience at school persuaded them to become teachers

> ...one of the lay teachers at my convent school who was my maths teacher for my O-level and then my A-level courses was really a great inspiration to me.... I look back to lessons all those years ago with her and I remember them all with great fondness and such magical memories I had really in my sixth form. It was very special and in some ways she

> had such a personality and sense of humour and I think that a lot of the way in which I teach is actually reflecting some of the practices I think she used herself (Teacher X, female, mathematics).
>
> My best teacher taught Biology. She would sit on a tall stool and simply ooze Biology - she gave me such a love for the subject that I just had to come into teaching (Teacher O, male, Science).

Youth work provided the incentive for some to become teachers and in three cases the prospect of a secure job was also a factor. The least experienced teacher was a mature entrant to the profession who decided after 18 years in engineering that he wanted to work with people rather than objects.

However, without exception, interviewees stated that the most attractive aspect of teaching as a career was working with young people. Frequently the initial answer to the question was a single word - 'Kids!' - and throughout most interviews the passion with which teachers spoke about children is significant.

> The big surprise to me is how much I enjoy being with kids. I thought I would be too impatient. I actually like them a great deal and find them really good fun (Teacher E, female, English).
>
> I just love working with young people - they make you laugh, they make you cry - it's just wonderful to see their faces when they grasp something for the first time (Teacher W, female, Mathematics).

The strong reputation of the school for pupil care and support and the unstinting dedication of the former headteacher in pursuit of pupils' welfare was recognised by OFSTED in the inspection of 1996 (Baxendale 1996:para 146) and was clear in the comments of many of the staff:

> Kids need to know that 99% of the time you are level - balanced; you don't scream and shout at them; they know you to be reasonable; they know you to be fair. Kids need to know that there's justice; they need to know they're safe; they need to know they're secure. They are really basic things. It's just about how you talk to people - it's a fundamental thing. Kids want role models; they want somebody to respect. Its clichéd I

> know but its - but that's effective teaching and its got nothing to do with numbers (Teacher Y, male, English/Mathematics).

Drawing on research mainly about elementary school teachers in Canada, Hargreaves (1994:141-157) noted that the 'commitment to goals of care and nurturance is a significant source of depressive guilt among teachers'. He acknowledged that 'a little guilt can be good for one' and that a 'strong care orientation, balanced with other goals...has been found to be strongly associated with positive school climates that in turn foster student achievement' - but teachers *can* care too much! Teacher Y was thus one of several staff that articulated some of the consequences of the strong caring image of this school and the enduring legacy of the former headteacher:

> I don't have time for target-setting - [Teacher G] produces the department targets on the computer and I just look them over and maybe alter a few. I certainly don't like sharing them with the kids - anything below a C is demotivating (Teacher N, male, Science).

> I feel that if you are asking me to do that [target-setting] then I'm going back to my engineering roots - treating kids as inanimate objects - and they're not (Teacher Z, male, Mathematics).

> [Discussing a low Science set] These kids had low expectations of themselves - "we're no good, that's why we're in here". I said "who says you're no good". They said "everybody does". [interviewee got emotional - upset] Something has gone wrong in education when kids in Year 10 see no relevance at all. I would throw the syllabus away and properly teach - it is more important that they should understand who they are. They get a negative image even before they come to this school - labelled. They need to be cared for as individuals - old JB [former headteacher] would understand this - not statistics as someone else sees them [reference to Deputy Head (curriculum)] (Teacher H, male, Science)

> I try my best to totally ignore it. I do what I have to with the target-setting...and I showed each individually. I watched their faces draw, I saw them sort of sigh.... I just think that - don't tell the kids, they don't need to know - have a go at me (Teacher R, female, Mathematics).

The emotion of working with, educating and supporting 'kids' clearly drives initiatives such as target setting and the use of value-added data down in teachers' lists of priorities. This was emphasised in various ways by several teachers. Teacher W was typical:

> I'm not going to upset my relationship with Year 10 by telling them that their targets are E, F or G. It's not rocket science anyway - we have always had to produce predicted grades. I'll hand them in but as far as the kids are concerned I'm their Maths teacher and that's the most important thing (Teacher W, female, Mathematics).

This is a finding that Hargreaves is unlikely to have encountered in his research over 7 years ago. However, it goes some way to explaining why the Deputy Head appointed to introduce target-setting (see chapter 3) had difficulties in getting staff co-operation. This will be discussed further in the next section.

A small number of teachers reflected on unpleasant childhood memories, for example in a rough school (teacher C) or being labelled as a failure as was teacher R:

> I left primary school illiterate, virtually. I must have gone to the worst primary school - its unbelievable. And when I went to secondary school they had remedial classes and I walked in, at 11, and in those days you went in this class and you never came out...somehow I'd had some sort of conversation with someone, my primary head or there'd been a conversation that I'd overheard, when they were categorising people, and I heard them say that I was the weakest of the...and I was really hurt and very upset. So I got to this secondary school and I thought I can't stay here, I really can't, and luckily enough for me I ended up with some superb teachers and I worked my butt off and I got out of there and I started working and I worked all the hours God sent .... I didn't think I'd be clever enough, but I would aim for that, I would aim high and for me being a teacher was flipping high...and if I made it I would try and make a difference to all the kids like myself that had to put up with such dross in education and that where I got it from. So I just kept going to see if I'd get there. And I did! Just! (Teacher R, female, mathematics)

This interviewee mainly teaches pupils with Special Educational Needs. She is one of the most outspoken teachers at Dalestone School and her views are often at odds with the Senior Management Team. She is, however, a teacher who is highly regarded by colleagues (including Management), pupils and parents, known for her effective discipline, organisation, and skills as a Form Tutor.

Only one of the interviewees revealed a critical incident that resulted in a significant change during her teaching life. A guidance and counselling course gave her the incentive to adopt a completely different style in which peer-tutoring plays an important part. She rearranged furniture and encouraged group work, movement around the room and discussion. This case simply illustrates that plethora of factors that get in the way of making valid comparisons between individuals and which were discussed in chapter 2. This teacher has adopted a style that she would be the first to admit might not work for other teachers. Further, she has had the opportunity of appropriate training which others have not received; she has been given the professional freedom to utilise this approach and has the benefit of a timetable which provides her own teaching room.

> I do like to close the door and my teaching room's mine. Nothing can interfere with me. Mine, and I know what I am doing.... I am comfortable here. I have got everything that I want. I work hard at making the room beautiful. It is a very comfortable classroom. (teacher T, female, Mathematics).

The notion of contentment will be of significance in effectiveness. Teacher L also made a strong mention of the importance of his laboratory in his teaching

life and like Teacher T he had personalised his work-space - even to the extent of painting it himself!

It is well known from the world of industry that a contented worker is a productive one. In contrast, several teachers pointed to issues that they were unhappy about and which they voiced as negatively affecting their practice:

> I'm a nomad - I teach in five different laboratories and on some occasions in classrooms because there are not enough labs. I've no place to call my own where I can mount displays or store resources (Teacher O, male, Science).
>
> This year I had a lot of afternoon lessons on my timetable and had a lot last thing in the afternoon including Friday. The few morning lessons were for investigative work but afternoon lessons became a very much more regimented - almost a textbook type of situation (Teacher X, female, Mathematics).
>
> [Teacher P] being off a lot - we tried our best to keep the kids on track.... We've had to go to lengths where the cover teacher has come in and sat with my class because they could do some written work and I've gone in and taught [P's] class (Teacher G, male, Science).

The majority of teachers revealed a passion for their subject as one of the most attractive aspects of their job.

> It's about dealing with literature all the time that is the main thing (teacher E, female, English).
>
> I had been in electronics but there was no spark [laughs at pun]....I wanted more out of my love of science then it suddenly dawned on me that I ought to teach (teacher G, male, Science).
>
> I have always liked numbers - there's a beauty in Maths you know and its great to see the sparkle in children's eyes when they see it too (teacher W, female, Mathematics).

Thus English staff usually spoke about literature and mathematicians about the beauty of numbers, whilst scientists would frequently recount a story of a particularly successful piece of experimental work. Teachers were, however, more varied in their responses to the National Curriculum or other initiatives

such as the National Literacy or Numeracy Strategies. Although there was a spectrum of responses, they generally fell into three categories in approximately equal proportions. Some staff, particularly English teachers, openly embraced the various reforms. This is probably a reflection of the leadership and passion provided by the Head of Faculty but is also likely to be related to the stability of this largely female team:

> I'm quite excited about the Literacy Strategy. I've seen the Literacy Hour in primary schools and have had some training. Its going to be hard work but children will benefit as their skills develop (Teacher D, female, English).
>
> With only being in teaching for 4 years, I've been with it all the time - I don't find it constraining (Teacher Z, male, Mathematics).
>
> Chemistry at GCSE was getting so watered down that I agreed with some of the kids that there wasn't any point because you'd get to an interesting bit then move on. Now it's the other way with a bit of challenge and a bit more interesting (Teacher G, male, Science).

Teacher M echoed the conclusions of Donnelly and Jenkins (2001:129-133) about the National Curriculum in Science:

> I hated the first edition with 17 Attainment Targets but once the changes had been made it was OK. For a while we seemed to do less practical work but I think we've cracked that one now (Teacher M, male, Science).

A second group of teachers were generally positive about changes but added notes of caution often expressing a lament for the loss of opportunity for them to be creative:

> It was clear 4 or 5 years ago that some of us were thinking 'skills' and some were thinking 'content'. So I have focused on skills for quite a long time but on the other hand one of the things that has always been a plus in the job has been the creative side and some of that has gone and that is negative.... I'll have to live with it even if it just takes some of my personal pleasure away (Teacher B, female, English).
>
> I'm quite looking forward to the Numeracy Strategy because I seriously believe that its needed - too many kids can't do simple mathematical operations without the aid of a calculator. I do worry though about the

> proposed delivery method - 10 minutes for this, 20 for that - what's going to give? (Teacher U, male, Mathematics).
>
> I think that there is still a healthy amount of autonomy within one's own classroom and even with prescriptive syllabi and schemes of work one can still have one's own teaching style and encourage a variety of learning styles among students (Teacher K, female, Science).

Several commented that a degree of flexibility was important for them as professionals to maintain interest, to permit them to explore elements of interest or to include something that they personally felt was important. This sentiment was eloquently expressed by one of the English teachers:

> The curriculum should be a coherent, cohesive structure - but should provide options and some element of flexibility within it. The analogy of cadenza within a concerto is pertinent, or a jazz improvisation within a 12-bar framework (Teacher F, male, English).

The third group of teachers was generally negative about teaching reforms. Their reasons mainly centred on their loss of autonomy and for some this led to reflections on the 'good old days':

> I feel it's [referring to National Curriculum] - I was going to say a Sword of Damocles hanging above you but I think that's going a bit far - but I'm always aware of it and that if I find something of interest or the kids find something of interest you are looking at the syllabus and saying 'but I can't spend too long on this' (Teacher L, male, Science).
>
> The A stream took Agricultural Science, the B stream did Biology and the Cs did Mode 3 CSE Rural Studies - it was relevant as many of the kids went into farming.... I feel frustrated by Double Award - there was no hassle in the classroom then (Teacher H, male, Science).
>
> Teaching is now so different. When I started [in primary education] we just followed our noses really - when the sun shone we went out and if it rained we didn't bother with a break. I've only glanced at it [the Literacy Strategy] but I don't like the idea of little bits of lessons being prescribed. I can see that it is about empowering children with their own language in a way in which our vaguer and sloppier teaching hasn't done but in another way everything becomes small and, I don't know, things that I really enjoy doing have gone - we're having a 'fun famine', we are rushed to get things done and we skid across the surface (Teacher A, female, English).
>
> It [teaching] *used* to be having a sense of autonomy. It *used* to be that you could, as long as things didn't become silly, to a reasonable extent you could do things your way and you could experiment.... The latest

> thing is numeracy, isn't it, which looks as if its going to be an edict...very prescriptive even down to the timing you spend on the introduction of your lesson to teach conversion of decimals into percentages.... The whole thing now seems to be sort of impregnated with an assumption that we're not doing the job properly (Teacher S, male, Mathematics).

Dalestone is thus fairly typical in that the introduction of new initiatives has passed through (or is still progressing through) typical stages as discussed by McCulloch *et al* (2000:75-6). Some staff are clearly unwilling or unable to change; like Teacher A, they might understand the rationale behind an initiative but are so overwhelmed by other, perhaps selfish, concerns that they do not move from a particular point of view. Others, who initially felt hostile to change, have reinterpreted the initiative and then (with growing familiarity, confidence and awareness of the 'spaces in which to exercise their professional judgement' (ibid. p.76)) have expressed acceptance and have even embraced the change.

> ...the National Curriculum told us what to teach. Once we had got used to that it was apparent that there was still room for manoeuvre in there (Teacher T, female, Mathematics).

Yet there are different kinds of acceptance. Teacher I, a scientist with a particular responsibility for SEN classes talked about 'determining the programme' she delivered and 'adjusting this as appropriate'. At the other extreme, Teacher J commented that he had been deprived of any opportunity to exercise his professional judgement and had acquiesced.

6.2 Effective Teaching

An important part of the interview schedule was to elicit teachers' views on what makes an effective teacher. Many researchers (for example Huberman 1993; Hopkins, Ainscow et al. 1994; Harris, Bennett et al. 1997) have produced lists

giving the characteristics of effective teachers and the Hay McBer report (2000) gives what is probably the most up-to-date model of teacher effectiveness. The latter was aimed at taking forward the proposals in the Green Paper (DfEE 1998) by giving detailed descriptions of teacher effectiveness based on evidence of what effective teachers do in practice at different stages in the profession. Hay McBer recognised three main factors within teachers' control that significantly influence pupil progress - teaching skills, professional characteristics (16 of these) and classroom climate. None of the teachers interviewed had read the Hay McBer Report and only a few had heard about it. However, the majority were aware of the headings on the Threshold Assessment application forms and that these corresponded to the government's view of characteristics of effective teaching.

However, in this analysis two of Fitz-Gibbon's notions are crucial. Firstly, we should not overstate the 'findings' in school effectiveness research (Fitz-Gibbon 1996:105) since many are based on 'correlation' that does not necessarily equate with 'causation'. Thus to simply match teacher professional characteristics with the performance of their classes would be unjust not least in view of the findings in the last chapter about the Critchlow-Rogers Effect. Secondly, 'what is educationally important'? (Fitz-Gibbon 1996:167) - in this case what features do teachers feel make them effective and which ones are most likely to impact on pupil progress as measured in terms of value-added.

Teachers were asked to explain how they taught various groups; what their expectations were; the pace of the lessons and the nature of their relationships

with class members. As far as was possible two contrasting groups were chosen with the aim of ascertaining whether teachers work differently with them and consequently obtain different value-added outcomes. There were indeed differences that will be discussed below and related to value-added data in chapter 7.

With notable exceptions upper ability sets were treated very differently to those containing less able pupils. In the classic pattern that has been identified by Jo Boaler *et al* (2000) with Mathematics sets, teaching in high sets is characterised by high expectations, fast pace and high pressure to succeed. Staff often talked about these sets in terms of high expected grades:

> I had very high expectations of them - my expectations were so high that they were aware of the fact that I was anticipating getting - and putting across to them the idea that we would be surpassing previous GCSE groups in terms of the final grades.... I did work them flat out (Teacher X, female, Mathematics).
>
> I set them very high expectations and challenged them continually (Teacher K, female, Science).
>
> We pushed ahead like an express. They knew that we were expecting everyone in the set to get A*, A or B grades (Teacher N, male, Science).

In contrast with low sets the expectations tended to be low, pace was slower and opportunities were limited. Teachers seem to expect less from these pupils in terms of homework and there was often repetition of subject material that has been taught at Key Stage 3. A focus on basic literacy and numeracy skills was also common. Many of the comments about these pupils were negative even where discipline did not seem to be a significant issue.

> Set 6 were a different proposition entirely. There were a few disaffected boys in this class and significantly I can remember all the boys' faces but only a couple of the girls....I tended to concentrate on strategies for

> passing the exam rather than the more widely philosophic approach which I adopted with set 2 (Teacher K, female, Science).
>
> ...what they were offered was pretty much a repetition in terms of content, the work that they had been doing all the way through (Teacher S, male, Mathematics).
>
> A low ability set but a pleasantly small group. Very contrasting personalities and my overall expectations were not great, although there were times when I was pleased with their efforts (Teacher J, male, Science).
>
> This was a very disappointing class. Tom broke an arm and never mentioned it. Jack was alright in the end. John was deeply disappointing.... Sally was a bit daffy and Peter was completely off the wall. Gemma was...em...she had her mind on other things. Kieron was a total pain. Sam was all mouth...(Teacher A, female, English).

Several staff felt the need to clarify the word 'expectations'. In particular there was a need to distinguish between *actual* and *relative* expectations. In most cases teachers commented that A* and A grades were high expectations of a top set but where D and E grades were expected of a lower ability group they were keen to emphasise that these were nevertheless relatively high given the ability profile in the set. Several staff, when pushed, admitted that they did not put as much effort into the teaching of the lower ability sets:

> [Discussing a lower set] I think you learn when you have got to the point when any more pushing and it will go the wrong way. [Question: but you are prepared to push harder with a top set?] Oh I push a lot harder with a top set (Teacher T, female, Mathematics).
>
> I was frankly not prepared to put in the same effort with set 5 - I didn't feel the need to. They were OK as far as behaviour was concerned but somewhat lacking in grey matter (Teacher N, male, Science).

Often teachers clearly did not want to admit to giving preferential treatment to the top sets. However, the language used strongly suggests that this was the case, and sometimes was not particularly covert. Frequently, references to top

sets were enthusiastic, the children were 'lovely', and 'we' was used to indicate the feeling that the class and teacher were together, a team:

> This was the first top set I had had here at Dalestone and that probably did to some extent actually, ultimately make me make the decision that I made [to push for top grades]. So I mean...I actually did it selfishly perhaps (Teacher X, female, Mathematics).
>
> The top set were like piranhas - toss them a few chunks and they'd devour it (Teacher H, male, Science).
>
> It was not so much the pace of the lesson but what we went off into. We were hitting areas of A-level, economics and sociology (Teacher H, male, Science).
>
> We had a lot of laughs and fun as I recall but there was a working atmosphere in the class and a competitive spirit (Teacher K, female, Science).

The reverse seemed true for many of the lower sets. References were to 'them' and generally more negative as has already been illustrated for Teacher A above. There was also a frequent tendency to offer 'excuses' for these classes and sometimes the language used implied that the teacher really felt that they had not given the lower set the best opportunities - that they felt guilty. Teaching strategies changed, often appropriately but some emphases, for example on 'discussion' work with these sets, suggest at least an element of capitulation by the teacher. From my position as Deputy I know that many of the problems expressed, such as discipline or absenteeism, were justifiable but when coupled with other sentiments I cannot escape from the feeling that some of these lower ability classes got a 'raw deal'.

> I remember being a bit apprehensive [about this class] because I saw one or two names and I think that's unfortunate...you can be influenced a little bit too quickly if you are not careful.... You perhaps think 'well, they are not really going to put in 100%...am I going to bother?' (Teacher X, female, Mathematics)
>
> [Question: what about your expectations of set 6] Realistic I think....it's my job to maximise the sunshine and encourage as much blossom as possible. But you can't make something that hasn't got buds blossom....

> You're pushing with the same amount of force [as with a top set] but with the less able kids you're pushing through a much more viscous medium and you're not going to get as far because their rate of learning is much less (Teacher S, male, Mathematics).
>
> [Set 2] were not consciously taught differently [to set 1]. Maybe it's just that more don't keep up in set 2 even though their potential is higher but they are in set 2 because of their performance lower down the school (Teacher B, female, English).
>
> A student teacher had that class for a while and they lost their impetus and never really recovered it (Teacher A, female, English).
>
> Set 6 needed a lot more encouragement and gentle handling because they thought they were 'thick' - their word - and unfortunately this was constantly reinforced in the exam system (Teacher K, female, Science).
>
> My set contained three with statements, three poor attenders who were monitored by the EWO [Education Welfare Officer], two who were persistently disruptive and one was pregnant (Teacher Y, male, English/Mathematics).

One teacher expressed a difference in his approach between top and lower sets in a different but equally poignant way. His statement refers to more able pupils in a way that implies that he does not have the same anxiety about less able children:

> Bright kids are almost a worry. You have got to get it right. "She is going to do well and if she doesn't I'm going to feel bad about that". It was almost like a worry - "I must make sure she gets an A*" (Teacher L, male, Science).

Two exceptions to the general pattern illustrated so far in this section need attention - these relate to the teachers of sets of pupils with Special Educational Needs and the majority of the English teachers. The SEN teachers stressed their high contact with their pupils:

> I have taught most of these pupils since Year 7 - we know and respect each other (teacher V, female, Mathematics).
>
> These groups of children benefit from having a single Science teacher with consistent expectations and a lot of personal/social knowledge of the group built up by regular contact. There is high teacher investment in the group because of day-to-day contact, a definite sense of ownership (Teacher I, female, Science).

Within the English Department the influence of the Head of Department is clear. There is a distinct contrast with her predecessor (Teacher F) and the majority of the staff were saying similar things. This Head of Department (Teacher C) has reorganised the setting arrangements so that there is less of a hierarchy and a better gender balance. There are effectively two bands of pupils giving sets with a wide ability range:

> I'm not a big fan of setting at all because I think that you tend to find that the lower groups have lower expectations of themselves and the teachers have lower expectations of the kids (Teacher C, female, English).

Further she has introduced a style within English teaching that is highly structured, confidence-building and geared to examination success. With one exception, teacher A, all of the current English teachers independently repeated this pedagogic strategy. The following quotations are all from Teacher C:

> I always talk to kids about learning to drive because that was the hardest thing that I ever had to do...and doing GCSE coursework is a bit like doing your driving test in that you can be a really good driver but unless you get your set pieces and show the examiner what they need to see you will fail the test.

> My view of differentiation is that you don't simplify things because ultimately you want them all to get to the same level - you just give the steps to get there. A good lad might be able to jump this high - now I can't jump that high but I can get high if I have a ladder. So my idea of differentiation is a scaffold...I tend to put in the supports.

> You could say that's bad English because it's teaching to a test. I don't think that I sacrifice putting across what I love about English to do that - I hope that they get both from what I do.

When teachers were asked for their definitions of effective teaching their responses fell into two main groups. A minority gave somewhat esoteric answers but most responded with statements of a more pragmatic nature. The former included:

> Effective teaching alters a pupil's perception of the world and allows them to act in an autonomous way (Teacher S, male, Mathematics).
>
> I've never really thought of defining it - my children could name the effective and ineffective teachers. Part of being an effective teacher is to try and develop their humanity - their ability to deal with other people (Teacher A, female, English).
>
> Effective teaching is when the kids walk out of the door at the end of the lesson and I know and I can see on their faces that they've learnt something and they feel better about the subject (Teacher L, male, Science).
>
> I would say that an effective teacher is able to read the signs, being able to look at the individuals, look at the class as a whole and make decisions on the best way to present materials and to approach things based on what the class have told you. [Its also about] how you present yourself to the kids...are you passionate about your subject. (Teacher G, male, Science).
>
> Effective teaching is the guided progression towards autonomy (Teacher F, male, English).

Some of these responses were subsequently elaborated in the more practical terms that the majority used, that is, 'this is what an effective teacher does or is'. The consensus would be that an effective teacher is an amalgam of different characteristics much as was recognised by Hay McBer. There were several recurring themes that were significantly used by some teachers and not by others. These themes included high expectations, confidence-building, being a good communicator, giving quality feedback, discipline, enjoyment and rewards.

> I think probably it's fundamentally your perspective on human life - how you treat people. Effective teachers are concerned with relationships with students and if there are positive relationships there, then the effectiveness tends to come afterwards. I don't mean just like being nice to kids because obviously that isn't all of it - you can be nice to kids and get nowhere - but having high expectations of them, sharing the expectations with kids so that they feel that they can achieve at a certain level.... Effective teachers have high expectations, a certain firmness, and are fair (Teacher C, female, English).
>
> [Effective teachers give] quality feedback. Pupils should be aware of their own work and how they are doing so that they can be given confidence (Teacher B, female, English).

> Having strict and consistent boundaries, high expectations...and rewards on offer for good, consistent work and behaviour (Teacher I, female, Science).

> An effective Maths teacher is somebody who has made those kids enjoy their Maths...who has got all that is possible from those children....Children like a benchmark, ground rules whatever you like to call it - they like to know, and they love praise - I must give out more merits and gold stars than anybody else in the department (Teacher T, female, Mathematics).

> An effective teacher is somebody who can communicate at the right level required for whoever the audience is. Far too many people do not communicate and it sounds like a liberal, woolly, airy-fairy kind of argument and I don't mean that (Teacher Y, male, English/Mathematics).

Once again, and in keeping with the 'caring' ethos of the school, the emphasis strongly lies in the area of good relationships between teacher and pupil. Only two teachers mentioned that an effective teacher gets good examination results. One teacher went as far as to say that what she wears was important:

> An effective teacher is lively, has got a bit of verve.... Kids often comment "we like your clothes or your earrings". You are asking young people to look at you for a long time and you owe it to them not to be boring or speak in a monotone (Teacher E, female, English).

The Mathematics department exhibits a wider range of teaching styles than any other in the school. It can boast expertise in group work, chalk-and-talk, peer-tutoring, the use of investigations and practical mathematics. Teacher V (now retired), for example was renowned for playing darts with her SEN sets as a way of developing their numeracy skills. Hence it is not surprising that two mathematics teachers mentioned being a reflective practitioner as being important in their effectiveness:

> I do it [reflect on practice] sometimes sitting in the bath. We've tried to have opportunities to take time in the Department but usually there isn't time to do it (Teacher X, female, Mathematics).

> I am very conscious that my teaching style is a behaviouristic one. If it wasn't for the fact of doing an MA I think that I'd be losing out on a lot. The biggest thing I've gained is what can we leave the kids thinking

> about? My style is now beginnings and endings of lessons....My kids also have two exercise books. [One] I call their driving instructions [and] they hand that in as well as their calculation book (Teacher Z, male, Mathematics).

Given the general pattern of teacher responses to questions about top and lower ability sets it is not surprising that the majority of teachers tended to avoid questions which asked with which classes they felt that they were most effective. Many admitted to feeling more effective with upper ability sets but were reluctant to say whether they were ineffective with other sets. Responses were commonly in accord with Teacher A:

> I don't know, I think that rather than being an effective teacher you sometimes have effective lessons don't you and sometimes you have a situation where you know that they have moved on and at other times I knew what I meant and they didn't (Teacher A, female, English).

The reasons for the ineffective lessons - predominantly associated with lower sets - were not unexpected. They included discipline, teacher workload, the teacher's state of health and pupil attendance. A few, however, were more open in their replies to this question:

> It's a basic truth - some teachers are better with some ability classes than others. It's a bit of a generalisation but I feel that I am good with the higher ability. It is the 'brick wall' kids I find hard to deal with - they are not disruptive but they truly just don't get it (teacher E, female, English).

> My expectation is that I will have more of an effect on the middle to upper ability range.... My perception is that I won't have such an effect on the kids that at the end of the day are going to perform pretty badly. That's probably a weakness and something that I'm aware of but if I'm honest that's how I see it (Teacher L, male, Science).

Two important responses bring this section to a close. There has been a justifiable tendency in this section to criticise the differing teacher behaviours towards sets of differing ability. It is, therefore, of interest to note the experience

of Teacher R who adopted the same approach with two different sets and found that it did not work:

> I went in and did the usual and, in fact, I was quite shocked because they got on with the work and I'm bored, walking round and I say "Do you understand?" "Yes." "Do you want any help?" "No." And if I do tell them off they all shut up - and they work in silence.... I'm not used to it and I must admit that I find it hard to adjust (Teacher R, female, Mathematics).

Throughout the interviews few staff were questioned about 'middle' sets - and effectively these only occur in Mathematics and Science. In the case of Science, the timetable arrangements mean that all staff teach a range of sets and thus choosing a contrasting pair was relatively easy. With Mathematics one interviewee has tended to have more 'middle' sets. As was clear from the Crtichlow-Rogers analyses in the last chapter (for example, Table 7), the middle sets contain pupils covering the widest range of ability:

> The middle sets can be the hardest because you haven't got a clear agenda. Although we say that this set will work towards the intermediate exam - remember there are three tiers in Maths - really some kids are bordering on the higher level or at the other, foundation, end. You've got the biggest range of ability in these sets. It might not look like that when you only look at their Maths results but that's not everything - some are in the top set for English. On the other hand they can be the best sets to teach - they don't tend to have a concentration of disruptive elements for example (Teacher U, male, Mathematics).

From a study based largely on the responses of pupils and classroom observations, Boaler *et al* (2000:646) concluded that ability grouping (in Mathematics) diminishes the 'opportunity to learn'. Further, they suggested that the traditional British concern that some of the able students succeed results in a majority achieving well below their potential. So far, this analysis of teacher responses from Dalestone School would seem to support those conclusions.

6.3 Performance-related pay

> *It is just possible that performance-related pay is a waste of public money. (Fitz-Gibbon 1996:195)*

Interviews were largely conducted during the period when teachers at Dalestone were considering, receiving training about, and making their applications for Threshold Assessment, and subsequently waiting for the External Assessor's visit to confirm or otherwise the Headteacher's judgement. This visit was relatively late compared to many schools and many successful teachers did not receive their threshold payments until June 2001. Thus the interview period was one during which performance related pay was high in the minds of colleagues.

.

The most significant observation from the teacher interviews is that there was clearly a high level of confusion and soul-searching going on in their minds. The majority of teachers in some way or other made the 'principled' statements along the lines of performance related pay being 'divisive' or 'bureaucratic' - no-one said anything more positive than 'I don't have any problems with it'. It was also clear that most of those eligible would also be making an application and would accept the money if successful - adding a caveat that they work hard, deserve it and could not work any harder.

> We totally don't want anything to do with pay in relation to pupil performance. Pay related to taking on extra duties et cetera yes, but not pupil performance (Teacher Z, male, Mathematics).

> Divisive and unworkable. It insults the collegiality of the profession. It is founded on a ludicrous, unjustified faith in the public examination system.... Too bureaucratic and distracting for SMT [Senior Management Team] (Teacher F, male, English).

> I feel uneasy about it and am not looking forward to making the application but if I'm successful I'm not going to say no on principle to the

> extra £2,000 - I think that's the way most staff feel - its really a long-overdue pay rise (Teacher N, male, Science).

Most teachers raised one or more ways in which they 'felt uneasy' about the Threshold process. The hottest issue nationally was the 'pupil progress' measure and the same was true for Dalestone staff. Staff expressed their concerns in different ways, some were, for example, critical of statistics, some were concerned for SEN colleagues and others were not happy about their pay being partly determined by what classes have been taught previously and by whom. Many mentioned the 'unmeasurable' facets of pupil progress such as growth in self-esteem. Some concerns were actually unfounded such as the teacher in the first example below who would have been perfectly justified in using his analysis in a threshold application:

> In Science we make use of the raw scores from Key Stage two and three don't we - and we convert these into fractions of levels. But only whole levels are reported. So a child moves from level 4.8 to 5.2, up 0.4 of a level in reality but he is reported to go from level 4 to 5. Another child moves from 4.0 to 4.9, up nearly a level but he remains a level 4. Which child has made the best progress?... I also don't like the idea of pupils determining teachers' pay, you know, how they perform on the day! (Teacher H, male, Science)

> I think its not appropriate to judge somebody on a group of kids that they have taught for one tenth or one eleventh of their time in school...is my pay really dependent on how much you can redress what has gone before? (Teacher S, male, Mathematics)

> I have no concerns about relating pupil performance to pay but I think that the notion of teamwork in teaching...is completely overlooked (Teacher K, female, Science).

> As an SEN teacher it is often difficult to measure attainment or performance in terms of NC levels or grades etc. Increase in self-esteem, self-confidence or organisation may be long-term outcomes which are difficult to measure (Teacher I, female, Science).

> Threshold is absolutely fraught with difficulties. There are lots of problems in terms of data collection, reliability of statistics [mentions CATs, NATs, YELLIS, Autumn Package], whatever standards you use in your analysis (Teacher C, female, English).

Potential abuse of the 'system' was another area of concern that was raised by several teachers. Teacher L's response was largely tongue-in-cheek for effect but clearly illustrated a concern that the process allowed teachers to ignore any poor pupil progress measures and only present their best results:

> If it suits me I'll cherry-pick and choose the ones that suit me and make me look good. I'm not going to pick up the ones that make me look bad - who would? And if you haven't got one to cherry-pick then hard luck! I want to do it for me. If everyone starts to do well then my groups are not going to look as good. I think its sending us in the wrong direction. I might want to hang on to an idea so I shine better than others (Teacher L, male, Science).

> Performance pay! [laughs] Largely, yes I'm happy with it as long as some teachers or Heads of Departments don't abuse it by giving themselves the 'cert' sets. I think that its easier to get good value-added scores with some sets rather than others (Teacher B, female, English).

A few staff made suggestions about how the money might be better spent, a better system of performance management, or offered 'fairer' ways of measuring pupil progress. The most constructive of these was from a mathematician who felt that looking at one year's worth of YELLIS data was unacceptable given the instability of the data. He proposed a minimum of three years' worth of data, perhaps looking at rolling averages. This might have suited one teacher who wanted to comment on the fact that she did not apply:

> I didn't apply because I got myself into a complete turmoil about it. I'd got the results from my Year 11s and looked at the statistics and decided that they had actually gone backwards with me. I burst into tears and said I can't do this. Some of the children in the class like [names notorious child] didn't help (Teacher A, female, English).

Teacher A was certainly not alone. Several Dalestone staff saw negative residuals and, not aware of, for example the Critchlow-Rogers Effect, at the time decided not to apply. Some subsequently changed their minds and sought ways to 'make the data look good'. Sadly, it was too much for some and three staff

were signed off by their doctors as a result of stress attributed to the Threshold process.

All of the 30 serving teachers at Dalestone who finally did apply were successful and are receiving their threshold payments. The Threshold Assessor who visited the school was impressed by the use of value-added data in applications. He confirmed the widely reported notion that his assessment would have a 'light touch'. Indeed, he commented that of the schools that he had visited very few had properly addressed the issue of pupil progress - yet most of those applications had been successful. Hence it was no surprise that three of my interviewees who had been critical of performance pay, came to me independently on Monday 16th July, 2001 to ask if I had read the front page of Friday's *Times Educational Supplement:*

> The Government's controversial multi-million-pound performance-related pay scheme has been a waste of money and has failed to affect classroom practice, new research suggests.... Heads who condemned the training they received to assess staff, found it hardest to judge pupil progress. (Dean 2001)

One commented simply - "told you so!"

6.4 Concluding comments

The Green Paper, *teachers: meeting the challenge of change (DfEE 1998)*, and more recent initiatives such as the Key Stage 3 National Strategy (DfEE 2001) plainly construct professionalism in a way that is some distance from that revealed by the interviewees. The government sees greater accountability, greater control over pedagogy and rewards for those who comply with their quest to raise standards. Teachers, though, tend to value their individuality and

the opportunity to be creative. They are aware of what they bring to the profession and particularly that their skills may produce outcomes in terms other than examination results.

There is an awareness of 'what works' and also most teachers are acutely aware of their own effectiveness even though they did not always express it verbally. There is also a group of teachers who are decidedly in favour of the sorts of reforms that, for example, the National Literacy Strategy will bring. Thus many of these teachers are not diametrically opposed to government policy – rather there are some areas of mismatch. Government policy does, however, unequivocally include rewards for good performance. The next chapter includes consideration of the extent to which value-added measures are of worth in this respect.

Chapter 7 Teachers on value-added

Teachers do make a difference, but I doubt that you'll find a mathematical way of judging performance. I'm reluctant to accept that value-added reflects teacher performance except perhaps at the extremes (Teacher H, male, Science).

7.1 Introduction - teachers' voices on value-added

This chapter considers the value-added results obtained by the classes of each teacher together with their interview data in an attempt to further address the research questions:

- To what extent can value-added measures be of worth in the assessment of teacher effectiveness?
- What are the effects of teacher attitudes and behaviour on the educational performance of pupils as measured in terms of value-added?
- What is the relationship between value-added policy and teacher attitudes and behaviour?

Saunders (2000:249) experienced difficulties in her analysis of interview data partly because interviewees confused different kinds of performance data. Also her interviewees often implicitly, rather than explicitly, expressed feelings about the proper place of data in their own and their colleagues' teaching. The initial analysis of the interview data for the present study presented similar difficulties. Often it was clear that teachers would go 'so far' in an overt way to express their feelings then slip into almost clandestine statements and actions (such as nods and winks) to avoid being 'unprofessional' in front of their Deputy Head.

In chapter 3 it was stated that Saunders' model to represent different 'emotional and intellectual stances towards the use of performance data' would be tested

against the teacher interview data. A summary of her definitions of these teacher responses is given in table 10 and these will be appraised in the following discussion.

Table 10 Teacher responses to performance data - a summary of Saunders' (2000:252) classification.

Response	Definition
Unengaged	Resistant to taking the initiative in making use of data; data is 'out there' and not intrinsically relevant to pedagogical needs.
Technicist	Enthusiastic reliance on performance data that is seen as the key to monitoring and evaluating pupils' performance. The meaning and interpretation of data is largely taken at face value.
Sceptical	Resistant to the literal use of data, though it is not rejected *per se*. May be concerned that the professionalism of teachers is eroded by the misapplication of figures.
Heuristic	Accept and value performance (especially value-added) data; the data is used for raising questions rather than making judgements. Crucially these are strategic questions about how well the curriculum and pedagogy meet pupil needs.

7.2 Teacher responses to value-added

All of the interviewees had at least a basic understanding of the principles of value-added and were familiar with YELLIS. Several teachers commented that they have baseline data for their classes in their mark books - indeed it is school policy that this should be so. However, the extent to which they were actively making use of value-added data or were enthusiastic about it varied considerably. As many of the quotations that follow demonstrate, teachers were often distinctly emotional when discussing value-added, being either hostile towards value-added or quite enthusiastic. It was thus quite easy to locate them on the 'cold - hot' axis of Saunders' model. However, her other axis, representing the 'intellectual' stance of teachers towards the use of performance data, proved to be difficult to utilise.

Saunders (ibid. p.250) takes the 'intellectual' axis to represent the degree of reliance placed by a teacher on data as a manifestation of pupils' ability in a given subject. It became clear that the difficulty in locating teachers on such an axis was largely as a result of the bunching of most of them at the 'provisional' end of the scale. No-one regarded value-added data as possessing intrinsic truth about the performance of a teacher or a pupil's ability. This is possibly a result of the training and debate that has occurred in the school over recent years about performance data. Nonetheless, it became apparent that this was also partly a negative reaction to the Deputy Head's voluminous publications and continuous promotion of performance data.

It is clear that many of the teachers regard the second Deputy Head as the only *technicist* in the school. Her enthusiasm for the data and perceived insistence that it is taken at face value and discussed at all levels is notorious:

> It was OK when we just had the GCSE results and YELLIS. Now [names second Deputy] gives us CATs, Key Stage 3 data for English and Maths as well as Science, YELLIS, those ridiculous Autumn Package graphs, and her own plusses and minuses against targets. It's paralysis by numbers, there's so much of it and it can become contradictory. I find that I can't make any use of it yet... [indicates second Deputy's office]... thinks that it's wonderful (Teacher N, male, Science).

The mass of data produced by the second Deputy and her apparent reluctance to accept 'excuses' creates this perception that she is a *technicist.* However, the truth is that her intentions are more about stimulating a debate about the performance of pupils leading to positive curriculum or pedagogical changes - it is, unfortunately, her façade that mitigates against this being appreciated by many teachers.

The introduction of target-setting and the school's Central Curriculum Record has placed the use of baseline tests and measures of pupil progress high on the school agenda in recent years and several teachers acknowledged this 'literal' use of data. This has become coupled with the 'accountability drive' of the second Deputy Head as has already been discussed in chapter 4. During interviews there were several negative references to the processes that were being used, 'self-fulfilling prophecies' and concerns that more harm was being done than good. Clearly this agenda has distorted teacher attitudes - several teachers were linking their negative opinions about the Deputy with value-added, for example:

> [Looking at standardised residuals for a Mathematics set] Look at these figures, they're all minus aren't they. Look at Mark with minus 2.1 and *she* [referring to Deputy] would say "what were you doing with Mark for all that time?" Well I was trying to motivate him - numbers don't take account of that. Mark was only interested in...boxing...but English and Maths were at the bottom of his list of priorities at 15 and 16 years of age. He was a young man. He was out on a Friday and Saturday night. He was doing the boxing and he was interested in lots of other things. That's not a cop-out. That's the way of the world, that was reality, and he comes out with minus 2.1. You could look at it another way - he *only* got minus 2.1, he *could* have got minus 5, he *could* have got nothing - depends how you look at it (Teacher Y, male, English/Mathematics).

Similarly, a former colleague acrimoniously regarded the YELLIS predicted grades as a departmental expectation of performance:

> ...above which we would reap institutional respect and status - but below which we would meet with overt or covert censure, humiliation, comparative disadvantage and diminished status (Teacher F, male, English).

It was this 'public' aspect of value-added data that was the greatest concern of a number of interviewees. Typical comments suggested that the data is unfairly used to 'bash' teachers:

> I'm not happy about the publication of so-called value-added that my set got last year. It had my name against it so I was made to feel small. What it didn't have was the fact that I shared the set with two other teachers,

> that the set contained idiots such as [names two pupils] and that I was off ill for a lot of the time (Teacher P, female, Science).

Situations such as these undoubtedly moved teachers towards the 'cold' side of Saunders' scale and it was amongst these that two out of the four of her 'responses' were most clearly represented - the *unengaged* and the *sceptical*.

Seven staff fitted with the description of *unengaged*. The former Head of English (Teacher F) was the most openly anti-YELLIS and distinctly resistant to the use of value-added data:

> [YELLIS is] a fiction or pseudo-science perpetuated by the institutional assessment of a predominantly non-verbal battery of psychometric tests on perplexed students - eminently inapplicable to English (Teacher F, male, English).

There was some evidence that these teachers were not confident in using YELLIS data or preferred to use other assessments such as Key Stage 2 or 3 results or reading ages. Of interest, four were Science teachers but probably the reason for their 'unengagement' is different in each case. Teacher J had recently retired and teacher N was close to retirement. Both indicated that they really had no interest in the school's preoccupation with statistics and that it undermined professionalism. Teacher I's interest was with children with Special Needs and she said that value-added data was not appropriate to her work.

> With SEN it is difficult to measure attainment in terms of levels et cetera. I am more interested in increase in self-confidence, self-esteem and organisation (Teacher I, female, Science).

Teachers P and Y also felt that the way in which data were used was unprofessional, that it was only something for 'management' and therefore they would not be involved with it. Of deeper importance is that they also had more

fundamental philosophical problems with value-added and reinforced Ball's point about 'single essentialising tags' (1997a:317).

> I wonder what value-added *really* tells you. You know, you talk about so-and-so as a number. That number doesn't take into account the host of things we know are relevant in education but you can't measure - what life is like at home, good days, bad days, stress, illness. I could go on but you get my drift....This data is really for the likes of you [points to me], for SMT or the governors if it is worth anything. It is meaningless for parents and the likes of me. I don't see how numbers are going to improve teaching, surely that's what professional development is all about (Teacher P, female, Science).

> I think that YELLIS is an invalid test. It is not a test of English, of ideas - there's nothing creative in it, there's nothing in prose. It's like an easy fill-in form. You give someone a form to fill in and make it easy for them and they'll do it, and they'll jump through all the hoops. You ask them to *write* something and its harder. I don't see a lot of linguistic capability in doing that YELLIS test. I don't like it and I think we get hung up on it. You can lay too much store by numbers and letters at the expense of knowing the kids (Teacher Y, male, English/Mathematics).

The essential problem with the *unengaged* category is that it sits uneasily on the 'cold' side of *technicist*, reinforcing the problems of the 'intellectual' axis, particularly at the 'literal' end. Thus the teachers who really wanted nothing to do with value-added data could hardly be described at the same time as showing a 'degree of reliance on data as a manifestation of pupils' ability' (Saunders 2000:250). Their only engagement with the data was basically when they had to, for example, with target-setting or reviewing examination results. Even then it was evident that there was a tendency to be disingenuous about the process:

> I find the whole process something of a navel-gazing exercise and, frankly, I pay lip service to it. It would be OK with one lot of numbers but there's too much (Teacher P, female, Science).

A key difference between the *unengaged* teachers and four that closely fitted the *sceptical* category, was in their emotional response to value-added rather than any intellectual difference. The former were 'colder' in that they were either

strongly against value-added or totally disinterested. On the other hand, the latter were less extreme and less likely to wholly reject value-added data. Hence, teacher A's comment, whilst having something in common with that of teacher Y above, is more moderate:

> I always treat it [YELLIS] with a certain amount of cynicism because I know what exams are like. Some of these tests like CATS and YELLIS are like sledgehammers to crack nuts - they are not necessarily closely related to the way in which children perform; they can sometimes give the wrong impression of how children cope with things for more than a minute at a time. These non-verbal tests have so many questions and so little time for each answer - tick, tick, tick - fill in the boxes and that is not the way we think in education so far as English goes. The whole experience is different so the test and the [GCSE] result are not at all related (Teacher A, female, English).

Also less openly negative about the use of value-added data, there were other staff who expressed some scepticism about value-added in that it did not reflect the 'caring' ethos of the school. Their comments have resonance with those of *unengaged* teachers P and Y above:

> ...they [value-added scores] take no account of what that kid has been through; there could have been a family break-up; there could have been a death in the family, there could have been numerous things....Therefore this child is an object that is going to produce something, that is going to give you a number (Teacher R, female, Mathematics).

> I don't like to see it in such a clinical way and to actually start picking up a bit of paper with just names facts and figures and then not appreciating perhaps in a wider context what that actually means.... I don't like the cold way in which we sometimes look at these things (Teacher X, female, Mathematics).

However, a vital difference probably emanated from their background in Mathematics. Neither rejected the data *per se*, accepting that 'statistics have their place in the scheme of things - I've got to say that 'cos I teach it at A level!' (Teacher X). Teacher R also strongly expressed the scepticism that she held.

She believed that several others were also unconvinced about the use of value-added data, particularly in relation to any assessments of their performance:

> Just 'cos you hold lots of figures towards me...isn't going to make me any better a teacher, it isn't going to make me try any harder, even with all this money palaver, that's not going to make me try harder (Teacher R, female, Mathematics).

In her analysis, Saunders did not have the complications of teacher accountability to consider. In her paper, there is only a brief mention of a Head of English using performance data to discreetly monitor the performance of teachers in his team (2000:253). Her interviewees, who were middle and senior staff in schools, were generally discussing the use of data in relation to the performance of pupils. In the Dalestone School teacher sample, however, teacher accountability was firmly on the agenda along with the use of data to meet pupils' needs. This difference meant that teachers were far more engaged with value-added in an emotional sense; Saunders' intellectual concept was distinctly lower in their minds, particularly those who have been discussed so far in this section.

Nevertheless, over half of the interviewees were positive about the use of data as a tool to provide information about children and, after examinations, for reflection on the progress that they made. Some credit for this must go to the second Deputy - her persistence has paid dividends since only senior staff were 'informed' four years ago. Some of the respondents admitted to using data in (as perceived by them) only a small way in their practice:

> I created the [class teaching] groups initially from the YELLIS information rather than the school information (Teacher Z, male, Mathematics).

> I've used value-added in recent years ...in making judgements about children - I go back to it at report time to make sure the reports are OK (Teacher B, female, English).
>
> [YELLIS is] just another weapon in the armoury to suggest how far you should be going, to the amount of challenge you give them (Teacher G, male, Science).

Two of the Heads of Department are beginning to use value-added data in a more strategic way for example in comparing sets or the residuals obtained by boys and girls:

> I use YELLIS personally to see how my kids have performed but also to see the performance of groups against each other - having parallel sets permits this (Teacher C, female, English).
>
> I've used the value-added data to modify the approach that I take in subsequent years. I also use it as a tool in differentiation. In the Department we've had an issue about the performance of boys so I use the value-added to see how they have done against the girls (Teacher H, male, Science).

There were few surprises, particularly for these teachers, when they were shown the value-added scores of the sets that they had taught. Although they may not have remembered the detail, most had seen the data at some stage. The pattern of top sets obtaining positive residuals was generally greeted with expressions of pleasure while the negative scores were generally received with comments such as 'I might have guessed'. The significance of the residuals frequently needed some explanation or revision, with several teachers seeing small negative numbers or zeros and thinking that they were 'bad' results.

This latter group, whilst by no means homogeneous, generally accepted and valued performance and the data with which it was announced. As such they broadly matched Saunders' *heuristic* category. It was the readiness with which these teachers talked about their classes, how they had taught them, ideas that

they had tried, and successes and failures, which particularly identified them. Once again, it was essentially an emotional response, characterised by willingness to talk about value-added tinged not with uncertainty or doubt, rather with a concern that data was used honestly and not in a narrow-minded way.

In addition to the complication of teacher accountability, there were three other issues that Saunders did not significantly encounter in her analysis. Several teachers, particularly teachers of Science, mentioned the problem of shared classes, and also of the teaching that pupils had received in previous years. The third issue, that of the results of individual pupils, was encouraged by the interview schedule and thus was addressed by all interviewees.

Science teachers were often reluctant to comment about the value-added of their classes and this was mainly because up to six or even seven teachers had shared them over the two GCSE years.

> I can't really claim that good result, neither should I take any flak for that bad one because I didn't teach those classes on my own. [Teacher N] only took them [lower ability group] in year 10, then [teacher J] did the Physics. [Teacher G] did the Chemistry and I did the Biology. If anything those poor scores for the lower sets come from having too many teachers - I think that we should only have two at the most (Teacher O, male, Science).

Many teachers commented that they were also unhappy (whether residuals were positive or negative) about taking any responsibility for results that derived from all of the teaching at Key Stage three as well. There was a contradictory sense of collegiality here when considered alongside comments about Threshold Assessment. On the one hand strong statements might be made about the immorality of a system that ignores the contribution of those who had

taught the children before. On the other a teacher might ask how that would be measured anyway, and then claim the residual as his or her own for the purpose of Threshold Assessment.

The greatest amount of fervent comment was, however, about the residuals for individual children. In the interview schedule teachers were encouraged to talk about their sets before seeing any residuals. Their comments after seeing the residuals were matched with those before and hence the schedule required teachers to talk about individuals. With very few exceptions the pupils that teachers had made comment about early in the interview were ones with residuals worth further discussion. This serves to illustrate the recurring theme that these teachers 'know' their classes, but also makes the point that any judgements need to take the 'exceptions' into account, such as exclusions, attendance problems and illness:

> Matthew was always doing well in Year 10 but he went right off the boil in Year 11....[on seeing the residuals] I was right about Matthew wasn't I! (Teacher G, male, Science).
>
> Craig was permanently excluded at the beginning of Year 11. He was eventually reinstated two weeks before the exam. He was predicted a C and got an F so its no wonder that he's minus 3 (Teacher O, male, Science).
>
> Alison stands out like a sore thumb. Really capable but half the time she wasn't in lessons (Teacher Z, male, Mathematics).
>
> I'm amazed, quite surprised that Adrian - I thought he'd be a minus. And that Daniel I was telling you about is plus 0.9! But some of them I'm a bit disappointed with them. You know when you look and you think [counts] 10 minuses! (Teacher R, female, Mathematics).
>
> Billy gave 100%, absolutely 100% all the way through....[sees residuals - Question: You mentioned Billy and he got the worst result of the boys...] Yes but this is now another issue itself. Billy worked hard in lessons but he didn't like exams. I'm even more disappointed now because I mentioned Billy.... Jody is another very hard working student, very keen,

> very enthusiastic...you put her into a test, in an exam and she goes to pieces (Teacher X, female, Mathematics).

The fact that the interview schedule required responses from all teachers about pupils has created a new level in this analysis. Whereas at a more superficial (whole school or even class) level Saunders' classification of teacher responses into four 'approaches' to value-added has a general application to the data from Dalestone School, it does not so readily at pupil level. Here no teacher could resist talking about individuals and in several cases there were apparent contradictions between the statements about individual pupil residuals and statements about value-added in a more general sense. Thus, *unengaged* teacher N remarked:

> I was pleased with Oliver's result – plus 1.25, that means that he got more than a grade above what he was predicted...and Helen's minus 0.8 wasn't unexpected, she had a lot of absence (teacher N, male, science).

It is important to note that although he commented and showed interest in the data, at no stage did he suggest that this in any way affected his teaching. However, most of those who at a general level could be described as *heuristic* mentioned something that indicated that they had thought about the implications of individual results for their work:

> [pointing to some of the negative residuals in a list] You know a lot of these are nice, well-motivated kids, but for all sorts of reasons they don't cope well with the work. In English they are often poorly organised so I've developed a set of writing frames that I give them as appropriate (Teacher D, female, English).

It was amongst the English teachers that this sort of response was most common. Saunders (2000: 253) found that her interviewees in each of the core departments were distributed fairly randomly over the four quadrants in her model. In this study although in none of the subject areas is any particular

response universally given, there is not a random distribution – the current, female-dominated English Department mainly fitting the *heuristic* category. Saunders' suggested that there may be a link between teacher responses and departmental cultures and this would be in keeping with that demonstrated by the English team; their common pedagogy and unity under the leadership of a passionate Head of Department being paramount.

The interview data revealed no one who would fit Saunders' *technicist* category. Anecdotal evidence and the views of teachers in respect of the Deputy Headteacher suggest that such a 'type' exists. One might speculate that their absence at Dalestone School partly reflects the ethos of the school and concern to meet pupils' needs. It might also be that would-be technicists are quiet and distancing themselves from an unpopular Deputy.

Irrespective of their overall leaning towards value-added, a majority of interviewees expressed some kind of anxiety about the use of performance data in the school. Often this was expressed in relation to the way in which it was presented to them or the 'paralysis by numbers' that was mentioned by one teacher. However, in spite of this, a majority demonstrate Saunders' *heuristic* approach, albeit frequently on the 'warm' rather than 'hot' part of the scale. The interview data serves to reinforce two of Saunders' conclusions (ibid. p.254); firstly, that although performance data may be neutral in intention it is rarely perceived as neutral in effect; and secondly, that the value judgements and complex technical issues behind the data are not always transparent or understood.

In testing Saunders' model against the teacher interview data a number of problems have been encountered. The most significant one is that her 'intellectual' axis seems inappropriate and mismatched with the quadrants on either side. Since the model was originally constructed for use with middle and senior managers it is possible that it simply does not have general application. Intriguingly though, the four 'types' that Saunders describes are useful in describing the responses of teachers to value-added. This has led me to consider that the model should be modified and constructed on a single axis.

Unengaged ------- Sceptical ------- Heuristic -------Technicist

This effectively separates the unengaged from the technicist to give a continuum of responses. The 'cold-hot', emotional response axis still applies and represents the enthusiasm that the teacher has for the data. This should be qualified to include the degree to which the teacher questions the data in a strategic way. Thus the *unengaged* are unenthusiastic and basically reject value-added data as being unsuitable for use in their work. *Sceptical* teachers are not rejecting the data as such but tend to draw back from the idea that they can inform their practice. The *heuristic* staff value statistical information and can believe that it has strategic worth in classroom practice. Finally, the *technicists* are enthusiastic about value-added data to the point that they do not question its merit and are blinkered to wider curricular or pedagogical considerations.

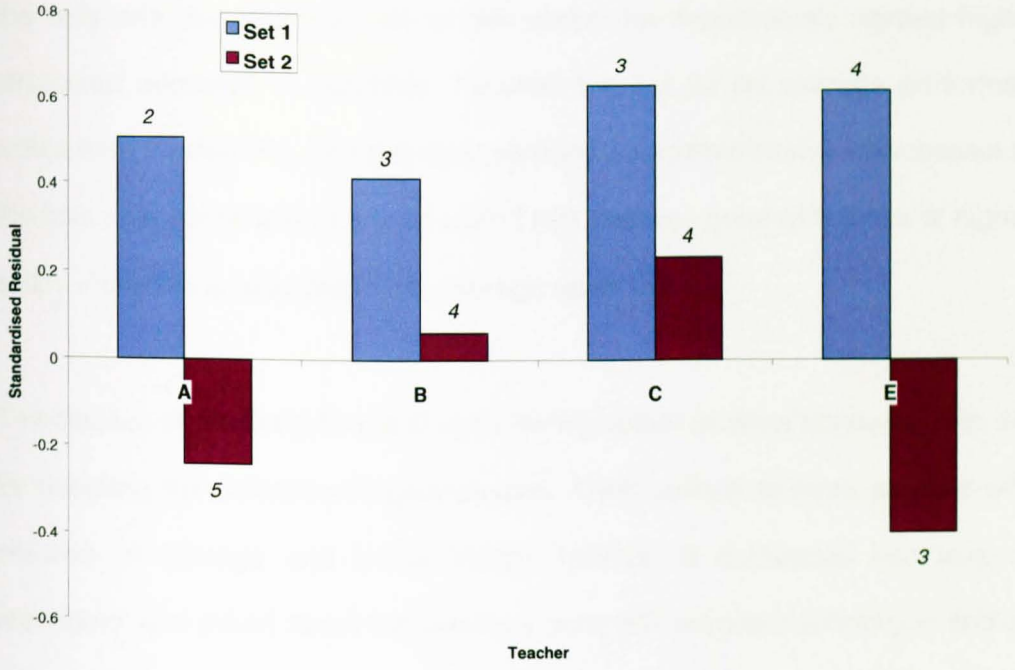
7.3 Matching value-added data to teachers' responses

In this section the discussion is mainly focused on four Mathematics and four English teachers. These are staff for whom the most complete sets of value-added data are available over the seven years. The sharing of sets in Science means that results cannot be attributed to the work of a single teacher and so these teachers have been excluded from this analysis. However, consideration will be given to some particular results, one of which will be from Science later in this section.

Figure 23 shows data for four experienced English Language teachers. They are all female and (with one exception) have served at the school over the period shown. This data is only incomplete in that Teacher C started at the school in 1994 and therefore picked up a class part-way through their course, and teacher E did not take one class through to GCSE. Across the seven years each teacher has generally taught top and second sets in rotation.

Figure 23 Standardised residuals for four English Language teachers

Data for four female teachers (A, B, C and E) are shown. All are experienced, specialist teachers. The figures are residuals (or average residuals where that level of set has been taught in more than one cohort) for each set taught by the teacher over a period of seven years. The italicised number at the head of each column gives the number of times that the teacher has taught that level of set.



In this analysis it must be acknowledged that the sample size is small and no account has been taken of the Critchlow-Rogers Effect mentioned in chapter 5. Despite this, teachers A and E tend to obtain negative residuals with second sets, whilst the other two obtain positive scores. All four obtain positive residuals with the top sets. However, the results at this level have been volatile over the years.

Teacher A has a different outlook to the rest of the full-time English staff. Although she has spent many years in secondary education she was originally trained for primary education. Further, she is the only full-time *sceptic* and her comments as reported in the last chapter ('Kieron was a total pain...') imply a negative attitude towards the less able and a lax approach in general. She was the only one, for example, not to talk about the department's agreed highly structured approach to teaching. Teacher E's set 2s on average performed worse than teacher A's. She too demonstrated a negative stance with respect to the less able ('brick wall kids') but stated that she was good with those of higher ability - and this is reflected in the average residuals.

The classes of teachers B and C consistently obtain positive residuals, with set 2s breaking the Critchlow-Rogers pattern. Their outlook is more positive with children of average and below ability. Teacher B expressed her 'love of teenagers' and talked about her previous work with pregnant schoolgirls and as a home tutor. Teacher C spoke of her own education in a rough school with low expectations and a loose regime, and like teacher R has a determination to make a difference. Teachers who have had difficult childhood experiences themselves, or who have worked with children who have disadvantages, seem to be associated with better-than-might-be-expected residuals with lower sets. Their feelings towards the less able are consistently positive and their understanding of learning needs is more acute. This group of teachers includes teacher I who has a background of work with EBD (Emotional and Behavioural Difficulties) children and therapeutic community teaching, and whose bottom Science sets always obtained positive residuals (average +0.3 compared to

minus 0.6 for the set above in the same years). Teacher Z who has worked in inner city youth clubs also belongs to this group.

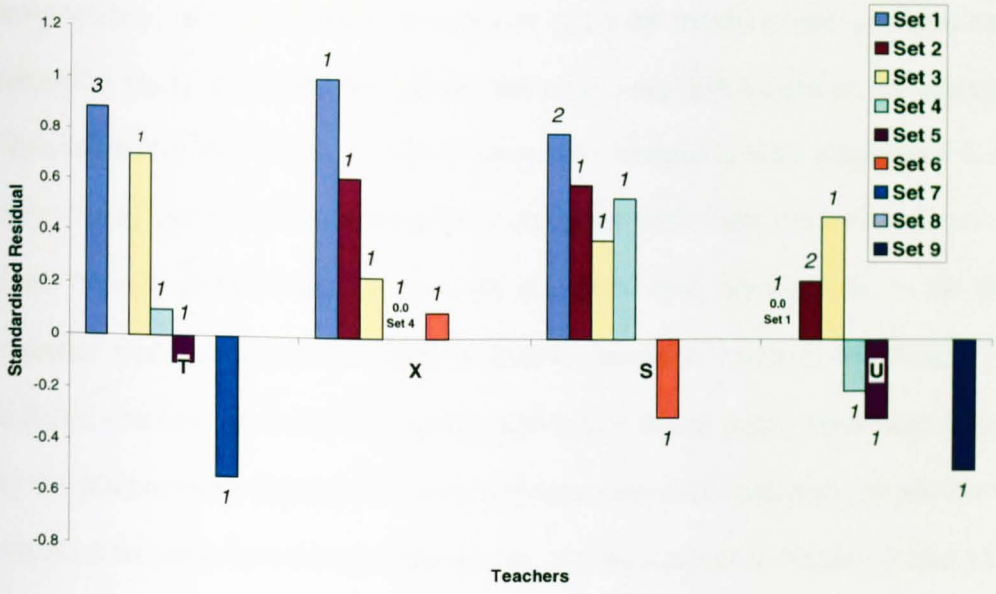
Thus, those teachers with more understanding and affinity with lower-attaining pupils tend to get better residuals from them. It is important to emphasise that this association does not automatically imply causation. However, the qualitative data strongly suggests that that the relationship is not coincidental. This evidence was to some extent reinforced by teacher L who suggested that one reason why he probably finds the lower sets more difficult to cope with was because he had a more privileged upbringing.

There are marked similarities between the Mathematics teachers and the English teachers in the inferences that may be made from the data. The residuals of four Mathematics staff are shown in figure 24.

Teacher T expressed similar attitudes to those of teacher E. He will 'push' top sets but 'you can't make something that hasn't got buds blossom' was his response when discussing the less able. Teacher S however, tends to have more success with those who find Mathematics more difficult. Her background in youth work, her father having been a Head of a Special School, a former role as a Head of Year and her unique (for Dalestone school) pedagogy that includes peer-tutoring, are pointers to this accomplishment. Teacher U's experience resembles that of English teacher A. His zero score with a top set suggests that something 'went wrong'. This long-serving teacher finds that he is generally content with teaching the 'middle' sets that he is allocated.

Figure 24 Standardised residuals for four Mathematics teachers

Data for four teachers (T, X, S and U) are shown. All are experienced, specialist teachers. The figures are residuals for each set taught by the teacher over a period of seven years. The italicised number at the head of each column gives the number of times that the teacher has taught that level of set.



It is important to appreciate that this evidence is largely of the sort that prompts questions rather than being conclusive in its own right. A succession of relatively poor residuals, or where upper ability sets return negative or zero residuals may raise questions of concern. However, such results may be due to chance or simply that a class has performed as predicted and thus scored zero value-added.

Alternatively, a particular set of results may appear poor against those of an outstanding teacher. Conversely, where performance occurs such as shown by

teachers B or C or where lower sets achieve residuals that are positive or close to zero there is justification for celebration.

In the attempt to measure the performance of teachers, a further note of caution is necessary, and this applies whether or not their teaching groups have been setted by ability. If a particular group had been very well taught in, for example, Mathematics in Key Stage 3, that group would effectively start Key Stage 4 in a very strong position in Mathematics compared with their performance in the more broadly based baseline test. All that their new teacher has to do is to maintain that advantage to score a positive average residual at GCSE. This focuses attention on the effectiveness with which these pupils have been taught by the Mathematics department over a longer period of time and not just on the contribution, important though this will be, of their teacher in Years 10 and 11. A number of teachers commented on this in relation to performance-related pay, for example teacher D:

> Performance pay for teachers doesn't take into account the contribution that those teachers who came before you made - or even contemporary colleagues - after all, the literacy work that I do supports all areas of the curriculum but I don't get a cut of your £2,000! (Teacher D, female, English)

In chapter 6 it was stated that teachers were keen to comment on the residuals obtained by individual children. When it came to a consideration of their performance teachers were keen to continue the discussion of individual pupils and, where there had been significant problems that probably adversely affected a result, suggested that some scores should be omitted from the calculation of a class average. I know from my work in helping teachers to complete Threshold Asessment applications (May 2000) that several did this. Teacher O, for

example, mentioned Craig who had been permanently excluded and another boy who had been a non-attender:

> The average residual for set 7 was minus 0.4. If you exclude those two it rises to minus 0.2 which is a lot better. I think that's fair but you need rules to ensure that people don't exclude anyone who, say, was absent for a term-time holiday and got a negative result (Teacher O, male, Science).

Teacher O's comment about fairness is relevant to this discussion. In an attempt to use class residuals to say something about the performance of a teacher the first hurdle encountered is the Critchlow-Rogers Effect. Where the residuals differ markedly from the 'expected' pattern or where the classes of a particular teacher consistently differ from those teaching parallel sets, then questions might be raised. The second hurdle relates to the individuals in a set and whether the results of specific children should be excluded. It is not surprising that most teachers want to exclude those contributing big negative scores but Teacher P offered another 'fairness' scenario:

> What if someone had a class where several children had private tuition for their Science and they all did a lot better than their predicted grades. The value-added then is not entirely down to the teacher (Teacher P, female, Science).

Teachers are seeking 'excuses' but at the same time are saying that they cannot take all of the credit. At a recent Subject Leaders meeting (Dalestone School, 2001, school records) concerns were expressed about the health of several able Year 11 pupils who were 'pushing themselves too hard'. Hence pupil motivation and the influence of parents are factors in the positive residuals obtained by some pupils.

It has been demonstrated that a simple statistical analysis of the value-added scores for a setted class has to be treated with caution, since pupils likely to

obtain positive residuals can be 'filtered' into upper sets and *vice versa* for those likely to obtain negative scores. The interviews with teachers introduced a multiplicity of factors which have the potential to make a difference to the educational performance of pupils, and which may be to a greater or lesser extent beyond the control of the teacher. Resources, the timetable, the learning environment, behaviour and attendance, setting policy and the composition of a set have all been cited as examples. Several factors are specifically teacher-related. Where sets are shared, as in Science, there are particular problems since the contribution of one teacher cannot be separated from another.

The teacher's attitude towards upper or lower sets is important; those who are positive towards the less able tending to obtain higher average residuals. The teacher's background frequently appears to be significant in this respect. There is some confirmation that pedagogical practices also make a difference. The evidence suggests that where many of these factors are present we can be confident that the teacher makes a difference to the educational performance of pupils in value-added terms. Thus teachers C and T have backgrounds that make for a positive attitude with all their pupils, they are concerned that their classrooms are attractive, and they adopt appropriate pedagogical practices.

The senior interviewee quoted at the start of this chapter captured the essence of these findings in relation to the research question about the use of data to measure teacher effectiveness. His view, which is substantiated by the numerical and interview data, is that the statistics cannot be used to judge the performance of teachers. Longitudinal data demonstrates Mortimore's

(1998:94) 'swings and roundabout' effect. Teacher A's top set in 1996, for example, obtained a poor residual (0.004) given the Critchlow-Rogers expectations of such sets. However, in 2001, she taught pupils in the equivalent class which, on average, obtained results over one grade higher than expected.

7.4 The influence of value-added on teachers' practice

Chapter 4 charted the evolution of value-added policy at Dalestone School since 1993. The cautious beginnings had little or no influence on teachers since the data were restricted to senior staff and even at that level it was simply information, it was not used to influence classroom practice. It was also true that the school had not found an effective way to manage the data and, at that time, the way in which the data was presented was not particularly illuminating. The Head of Mathematics was the first to use YELLIS data in a diagnostic way, but this was in fine-tuning setting arrangements, and again did not have any direct impact on teaching and learning.

The national target-setting agenda from 1996 onwards, though embraced by the LEA and discussed within the school, was effectively a paper exercise until the appointment of a new Headteacher and second Deputy. In a conversation with the new Headteacher it became clear that the Deputy was specifically appointed to undertake the difficult task of implementing a regime based on accountability. Repeatedly in interviews it has been clear that the notion of accountability and the way in which it has been promoted have been deeply unpopular. This unpopularity remains in the minds of several teachers and was evidenced by the

frequent use of 'her' in interviews. Yet this strategy has overcome the previous inertia and accords with Saunders' (2000:255) conclusion that value-added analyses are interventions that need to be managed.

Consequently, there are now distinct signs of acceptance and particularly that value-added data are being used in a direct way to inform and direct classroom practice. The coincidence of Threshold Assessment has lubricated this process. Threshold Assessment demanded that teachers reflect on their practice and provide evidence of pupil progress. That this has boosted the level of understanding about performance data was clear in the interview data; with some exceptions, teachers could discuss residuals and indicate that they understood what they meant. Rapid analysis and publication of performance data is now the norm, and the development of a school-wide computer network carrying the Central Curriculum Record has not only given every teacher full access to the data, but reinforced the expectation that they should access it.

The critical Ofsted inspection of December 2000 (Drew 2000) and the subsequent Action Plan (Dalestone School, 2001, school records) have also contributed towards the change in attitudes. With a demand for greater monitoring and evaluation coupled with a statement that assessment and target-setting need to be used in a more effective and consistent manner, coming from outside of the school rather than the SMT, some of the resistance has dissolved. The finding that about half of the interviewees fall into Saunders' *heuristic* category is evidence of this change. Thus Dalestone School is significantly in advance of the schools studied by Saunders and Rudd (1999:13)

where 'active use of value-added data turned out to be quite limited'. This is partly a reflection of the work of the SMT but also illustrates the changes in the national policy agenda over the two years since Saunders and Rudd completed their study.

Of greater significance is the impact that value-added measures are beginning to have on classroom practice. Several teachers commented that baseline measures help them to build a more detailed picture of each child that they teach:

> I think value-added, used properly, gives you a clear picture of where they're at (Teacher G, male, Science).

Teachers also said that they use data in forming groups within classes and that data informs the process of differentiation. Dalestone teachers are encouraged (and Threshold and Performance Management reinforce this) to reflect on the results of the last cohort. Prompts from the Deputy (Deputy Head, 2001, school records) focus discussion of pupil data on teaching and learning issues. Several teachers clearly use the data to inform the way in which they will work with the next cohort. Teacher S had looked at the residuals for one set on the basis of gender:

> I thought that the boys were a problem but maybe what happened is that the boys demanded attention, so the girls weren't driven along as well as they should be. I'm aware of that now (Teacher S, male, Mathematics).

However, concerns are being raised about the volume and variety of data and that it can be contradictory and 'clinical' - it was described by teacher N as 'paralysis by analysis'. There is some confusion about what is value-added data and what is something else, and differences in the YELLIS, the Autumn

Package, Key Stage 2 to 3, the second Deputy's own and the LEA's relative value-added systems have led to this confusion. This can lead to inertia or to busy teachers becoming reluctant to participate in its use:

> I'm no longer sure about the link between numerical evidence where they are now and extrapolating that – because that's what's happening and an extrapolation is dangerous – far more risks involved in projecting beyond, rather than between known data (Teacher S, male, Mathematics).

> I really don't have the time to take it all in. I suppose I don't do it justice – I ought to do more with the results but I find it baffling at times (Teacher D, female, English).

However, another issue is beginning to emerge as teachers gain proficiency in handling performance data:

> I know that the point is that we use value-added to look forward but I need help with that. I'm really disappointed with that set's results but what do I do to make sure that it doesn't happen again - I mean really do, not tinkering at the edges. I can't work any harder. It's like I feel I need to fundamentally change something about what I do - but what? That's the frustrating bit (Teacher X, female, Mathematics).

Teachers might have that more detailed picture about how their classes have performed in the past and appreciate some superficial ways in which they might use this information to inform their future practice - but at this stage they discover that they have professional development needs. This is significant for the managers of the school since, if value-added is ultimately about raising achievement, it is vital that they also put in place the systems and structures needed by professionals to make the best possible use of it.

7.5 Teacher influence on value-added policy

Teachers generally made little comment about value-added policy at a national level. The former Senior Teacher confirmed one of the conclusions from chapter

4 that by participating in schemes such as YELLIS, schools had influenced political thinking.

> When I was at one of the YELLIS meetings there was almost a revolutionary feeling that value-added was the way to beat the league tables (former Senior Teacher, 24-06-00, Interview record).

Most teachers, however, felt that they did not have influence at an elevated policy level as it was too distant from them. Performance data was not given to most teachers until 1998; the policy being that it was essentially for the information of managers – little thought was given to using such data in a proactive way. Subsequently, everyone was given data and for the majority there was a feeling that they were being 'acted upon' and that did not have much say in its use. This was inevitably closely linked with teachers' feelings about the Deputy Head that have frequently been raised in this study.

Colleagues were thus in a similar position to that experienced by teachers when the National Curriculum was imposed. Helsby (1999:166) noted the somewhat negative tenor of teachers' views that were associated with something imposed upon them and with which they were unfamiliar. Teachers went into a reactive mode until they found the 'spaces for manoeuvre' (ibid, p.169). With value-added at Dalestone School teachers seem be at this stage although their reactions are largely a response to the management of the process rather than to the concept of value-added itself. Thus there are confusions arising from the volume and variety of data, retorts to the expectations that are being placed on teachers, and concerns that professionalism and autonomy are being replaced by accountability.

Nevertheless, a number of teachers are becoming more proactive as was evidenced by the number taking a *heuristic* stance towards value-added data. They are 'coming round' to the view that there is value in a detailed analysis of data in the quest to raise achievement. However, this has certainly been mediated by Threshold Assessment that has encouraged teachers to reflect upon their practice and analyse the data pertaining to their classes. Threshold Assessment has also encouraged teachers to develop the skills necessary to question data. As confidence has grown teachers are seeing the spaces for manoeuvre.

In their examination of policy into practice and practice into policy in the primary classroom, Osborn *et al* (1997:52-64) identified four strategies of creative mediation by which teachers could be seen to have formulated classroom policy. One of these, conspiratorial mediation, has some relevance for this study but there are signs that the other strategies are developing.

Conspiratorial mediation should not be read in a strictly pejorative sense. There is no evidence that teachers are being deliberately subversive. However, as has already been discussed, some teachers are manipulating data in a relatively small way to produce a 'fairer' picture such as omitting the residuals of long-term absentees. Whilst this may be 'fairer', there is no consistent approach being taken and so some could easily manipulate data more than others. Staff were given the Autumn Package value-added graphs on a training day in September 2001. They had to complete these with the results for their 2001 cohort GCSE classes. These were collected and then copied to all senior staff

and the Governors' Curriculum Committee. Although they were not asked to, several staff also entered the names of pupils who were below the lower quartile line in order to 'justify' the poor result.

Finding this room for manoeuvre has been described by Goldstein (2001:439) as 'playing the system'. He has noted this in respect of 'league tables' and suggested that results would therefore, at least in part, reflect the ability of schools to manipulate the system, rather than reflecting the quality of education. Scott (1996:139) has described this as a 'gap between intention and implementation which allows schools and teachers to behave in ways not prescribed by policymakers. In the same way, results in Threshold Assessment applications may, in some cases, represent a teacher's ability to manipulate results rather than their performance as a teacher.

In the GCSE Subject Reviews (Dalestone School 2000, 2001, school records), there is evidence that some Heads or Department are making use of the data to press home a case for additional resources or staffing. Undoubtedly, this is a legitimate thing for them to do but value-added evidence does not usually back up their case. Hence, some responses to the question in 2001 'Do the value-added graphs for each teaching group indicate any resource, groupings or professional development issues?' include:

> Better-equipped rooms with updated cookers and hygienic work surfaces in rooms A8 and B1 are required. (Food Technology).
>
> How do we want to address the laddish culture? (Resistant Materials).
>
> Resource issue – rooms needed for dedicated theory work (PE).

Five subject areas did not complete this section of the review form and many left other boxes incomplete. This is not the usual response of middle managers but indicates that they are gaining the confidence to be selective. Science staff, for example, specifically stated that they wanted to wait for the YELLIS analysis before they would comment in full since that was the data with which they felt most comfortable.

'Protective mediation' (Osborn, Croll et al. 1997:57) is evident in a few teachers' reluctance to discuss targets (which are derived from prior attainment and value-added information) with pupils, particularly those in lower sets:

> I don't feel happy about target-setting. I know we're supposed to discuss targets with kids but I don't with the less able. I find it too demotivating for them – 'your target is an F'. If you say that to some kids their response is 'forget it'. I find that the chances graphs are better (Teacher X, female, Mathematics).

One teacher could be described as being 'innovative' (ibid p.60). Teacher G has developed an 'electronic markbook' with which he is 'continually comparing with value- added':

> It's always comparing, and I'm looking to see whether, on my scale of things, the kids are on target and do they match up to YELLIS or whatever data I'm using (Teacher G, male, Science).

Teacher G has also been influential in terms of 'collaborative mediation'. He has used his information technology skills to make value-added data more accessible to Science staff, for example in using Autumn Package and prior attainment data to produce lists of predicted grades to help colleagues in the target-setting process. This has been popular in that it has saved teachers a lot of time working out which target is most appropriate for each pupil. His work has

helped to make what many find unpalatable into a more acceptable part of their work. It can also be argued that the Threshold Assessment application process has brought increased collegiality and has consequently enhanced teacher professionalism.

7.6 Concluding comments

One of the most striking aspects of the teacher interview data was a persistent sub-text about teachers as professionals. Although this was evident in a number of remarks, non-verbal responses were particularly potent. I was both surprised and unprepared for the laughter, tears, passion and acrimony that reflected how many colleagues felt about value-added – I had expected many to regard it as a somewhat lacklustre subject.

When discussing value-added, teachers sometimes commented directly about their professionalism. This was usually in relation to feelings about a loss of autonomy and increased accountability. Several teachers, for example, saw the Green Paper as a means of redefining teacher professionalism - which was exactly its intention:

> I recognise that we are proposing a significant change for the profession, but I believe passionately that it is the right change (David Blunkett, Former Secretary of State for Education and Employment DfEE 1998:5).

However, many teachers in the sample did not see the proposals in the same positive way. They felt that their control was being eroded and this was especially manifest in discussions about Threshold Assessment. The evidence in the present study confirms a finding of Helsby (1995:324-5) about the effects

of the National Curriculum upon teachers' sense of professionalism. She referred to the notion of 'professional confidence' and found that when confidence was low, teachers were likely to take a more passive role and were amenable to manipulation and being 'told what to do'. At Dalestone School, teachers have generally complied, particularly when faced with an authoritarian Deputy who has insisted on the regular and detailed use of performance data by every teacher.

There are signs of a change at Dalestone School with its origins at Departmental level. The English Department is a clear example of the sort of relationship between departmental culture and departmental performance that was investigated by Busher and Harris (2000). Just as departments can be differentially effective, so they can be the source of school improvement and 'go beyond the strings of numbers to understand the trends being suggested and think about the implications for [the] curriculum' (ibid. p.71).

Chapter 8 Conclusions: the value of value-added

> *Teachers are better off using YELLIS than subjective evidence - if nothing else staff are very modest and poor about writing about their achievements - they're very self-deprecating; they don't blow their trumpets easily. Some objective evidence is better than subjective evidence. So it has a role and they ought to get better at using it and less threatened by it....It is the best indicator that we've got so far (Former Headteacher 16-08-00, Interview record).*

8.1 The framework for this thesis

'Teachers make a difference' is fast replacing the slogan 'schools make a difference' and this case study has sought to examine the extent to which the former statement can be justified using value-added data. A history of value-added policy at national and school level has been studied and this has been followed through to establish its impact on individual teachers. In a reciprocal analysis the effect of teachers on value-added policy and its implementation has been examined.

This study has responded to the call by some school effectiveness researchers to work below the level of the whole school. It adds to the newly emerging body of knowledge about departmental change and improvement. Notably, it engages at the level of individual classes and their teachers. It has, as a consequence, uniquely entered 'territory' that many have been reluctant to engage with. Thus value-added data, which are typically analysed at subject level, have been disaggregated to reveal the performance of individual classes. Another distinctive feature of this study is its longitudinal nature - that is, covering seven years.

The case study methodology has combined quantitative and qualitative techniques. Quantitative techniques have concentrated on the analysis of data for core subjects that have been provided to the school from the YELLIS project. This has been considered alongside the voices of teachers of the classes represented by quantitative data as recorded in semi-structured interviews. Triangulation was provided through a range of documentary sources and supplementary interviews.

A strategic feature of the study has been the researcher's positionality. Being a Deputy Headteacher in the case study school and also a parent has provided a special perspective on the data set. It is claimed that this has particularly provided 'entry into those innermost arenas and confidences' (Woods 1996:38) and thus a deeper layer of reality has been revealed.

8.2 Reflections on the research process

From the outset, this study has sought to make statements of educational importance rather than statistical significance. This has been because sophisticated techniques, such as multilevel modelling, have not only been shown to add little to that provided by, for example, simple regression techniques but also would be difficult for the majority of teachers to understand. Further, since that data has been analysed at class level, it is difficult to attribute significance where data from small numbers of pupils are involved.

The disaggregation of value-added data into classes has been a straightforward and successful process. It has mirrored the approach taken by many teachers when preparing applications for Threshold Assessment. As a result of using this method serious flaws in the common interpretation of value-added data for setted classes have been highlighted.

Teacher interviews generated richness of information that was beyond expectation. Transcripts revealed a stream of consciousness with long, free-flowing sentences juxtaposed with short, snappy ones. The interview data is thus compelling; teachers revealed their innermost feelings and passions and hence their 'voices' were clearly heard.

The period during which the interviews were conducted was longer than originally planned. This was partly due to pressures of work both for me and also some of the interviewees. Illness amongst some interviewees also caused some delays. However, they were held within the period over which round 1 of Threshold Assessment took place.

One of the most important aspects of the methodology has been the need to consider my personal influences on the research process. My background as a scientist was influential in the selection of the research area – I have a liking for, and a desire to explain, numerical data. The qualitative elements of the research were initially problematical, but with experience I have not only grown to respect and value the results, but have been changed by them. For example my awareness of the worth of symbolic interactionism as an ontological and

epistemological base has developed, and I more fully appreciate the micropolitics of schools and the values and attitudes that teachers bring to the classroom.

The Doctor of Education programme, and the research for this thesis in particular, has fostered my role as a 'researching professional' (Bourner, Bowden et al. 2001:71). The work has influenced my professional role - it has introduced new problems and opportunities. It has affected my professional positionality. I expected my role as a researching professional, specifically the most senior member of staff at the school next to the Headteacher, to be permeated with ethical dilemmas – and it was. Although, as chapter 2 records, I was well prepared, there were some issues that emerged on a scale that was greater or deeper than expected. So, where 'teachers' voices have frequently been silenced [or] repeatedly ignored or excluded' (Hargreaves 1996:12) they fully took the opportunity to express their feelings and emotions with me. This has provided a profundity to the research but also produced difficulties in the interpretation of a large volume of interview data. Inevitably something is lost in this process.

Another difficulty has been in dealing with comments, often of a negative nature, about my colleague, the second Deputy Headteacher. There have been many occasions where colleagues reached past what might be described as acceptable criticism, and at which time I might have interrupted the interview. With considerable difficulty at times, I continued the interview and in the final analysis believe that this was right to do. This is because it became clear that

these comments were significant in the interpretation of teacher responses to value-added data. The Deputy has effectively provided a focus for and clarity to teacher opinions.

8.3 From policy to classroom practice

In a climate of demands for greater accountability in the public sector, performance data for schools form the basis for 'league tables', OFSTED inspections and, more recently, Performance Management. The research community has demonstrated that value-added data can be more informative than 'raw' comparisons since they take into account the prior achievements of pupils. Many schools have paid for value-added measures such as YELLIS and have sent powerful messages to government about its worth. Yet government policy has consistently avoided genuine value-added measures, preferring 'benchmark' information and less-than-adequate Autumn Package 'value-added' graphs.

Whereas government have shied away from the use of value-added data in school comparisons, they have controversially included a 'pupil progress' element in arrangements for monitoring the performance of teachers (Performance Management) and in applications for Threshold Assessment. This 'progress' measure is to take into account pupils' prior achievement and is hence unequivocally a value-added measurement.

At Dalestone School, value-added has a history covering approximately eight years. During the first four years, YELLIS data was used by middle and senior managers essentially for the purpose of reviewing 'how well we had done'. Active use to inform future practice was rare and this pattern is in keeping with the findings of Saunders and Rudd (1999) and Saunders (2000).

The last few years have been characterised by the implementation of the national target-setting agenda coupled with changes in the composition of the Senior Management Team at the school. Now a wide range of data, including value-added data, has been actively disseminated to all teachers. Many teachers have reacted to this in a negative way. In part this has been a response to the data itself, but it has also been due to the way in which data and its use has been promoted. There is, nonetheless, a growing acceptance that performance data is now a part of the culture of the school and there are signs that teachers are using data to inform practice, albeit in small ways at present.

Teachers are not simply processing performance data in a mechanical way. They have agency of their own; they have ideas and expectations and are becoming significant in mediating the cultural changes that have occurred (and are occurring) particularly from a time that has been described by Ball (1990:153) as 'professional/collegial' to the current 'managerial/bureaucratic' (one might add 'accountable') regime. In a data-rich context, teachers as individuals and as groups are in a process of redefining their existence and the policy environment.

8.4 'Single essentialising tags' - a 'fairer' way?

The 'pupil progress' element of the Threshold Assessment process created the most unease for teachers. This was out of proportion when compared to the other standards – and this was in a relatively data-rich school. Most accepted that value-added data was required and the anxiety was rooted, even in the most *heuristic* colleagues, in a distorted notion that their professionalism was being reduced to a number. This research has demonstrated that although value-added is fairer than 'raw' data as a representation of pupil progress, with setted groups there are also problems.

The Critchlow-Rogers Effect (whereby residuals for upper sets tend to be positive and those for lower sets negative) has been a significant finding from this study. It has been described as a statistical artefact resulting from:

- ◊ A baseline test score (YELLIS, CAT or Average KS3 for example) being used to predict an outcome.
- ◊ Setting being organised principally on the basis of achievement demonstrated in that subject e.g. KS3 Mathematics score – that is, *not* the baseline.
- ◊ The outcome is a single subject score that is compared to the predicted score to determine a value-added residual.
- ◊ Setting on the basis of a subject score effectively filtering those pupils with an aptitude for a subject into a higher set than might be suggested by their baseline scores; and the converse for those with a lower aptitude for a subject.

It has consequently been shown in this study that judgement of the competence of a teacher on the basis of this data is seriously flawed. This applies whether the baseline is YELLIS, average KS3 or CAT data. Only when national subject-specific value-added data are available will the progress made in setted groups be judged with greater confidence. A significant conclusion from this research, therefore, is that value-added should not be used as a basis for performance-

related pay in setted situations. And yet we have a government that advocates setting, *and* wants performance-related pay. This is an issue that warrants a significant amount of further exploration by teachers, their representatives and the government.

However, there is strong evidence that the Effect is not entirely a statistical remnant. Teacher interview data reveals that many teachers have positive attitudes towards pupils in top sets, and are negative towards those in bottom ones. Strong encouragement of pupils will undoubtedly contribute to the movement of residuals in a positive direction whilst a lack of interest in or fear or dislike of pupils will do the reverse. Wiliam and Bartholomew (2001) and Bartholomew (2001) have recently presented evidence that accords with these findings.

One of the issues arising from the methodology has been the need for subject-specific baseline data. These are not yet available nationally. Using Key Stage 3 Mathematics scores for a group of schools as a baseline for a study of the influence of ability grouping, Wiliam and Bartholomew (2001) found that pupils in top sets obtained postitve residuals and those in lower sets obtained negative scores.

The literature search revealed several examples where data has been analysed and a pattern of high value-added scores for top sets and apparent underachievement by lower sets has been found. Such conclusions as have been drawn may, at least in part, be invalidated by the Critchlow-Rogers Effect.

Thus Saunders and Rudd's (1999:25) scatterplot data for mathematics in one of their case study schools indicated good achievement by top sets. This was contrary to an OFSTED team's suggestion that there was underachievement with these sets. By not being aware of the Effect, It is possible that Saunders and Rudd were misled by their data and OFSTED were right. Similarly, findings by Goldstein (1997a) and Kilyon *et al* (1997:10) that schools were being differentially effective in relation to achievement at GCSE may be unsound.

There are distinct exceptions where the arrangement of residuals across setted groups is different to the Critchlow-Rogers pattern. These centre on four themes – teacher background, pedagogy, shared groups and teacher effectiveness.

Where teachers have more understanding and affinity with lower-attaining pupils they tend to get better-than-predicted residuals from them. These teachers are predominantly those whose upbringing included some disadvantage, or who have significant experience of Youth Work or work with children who would be defined by DfEE *Circular 10/99 Social Inclusion: Pupil Support* (1999b) as 'at risk'.

Pedagogy can vary in different ways for example, from subject to subject, set to set and teacher to teacher. Whilst some pedagogies may be less effective, there is evidence, for example from mathematics, that very different styles can produce equally good results. Where lessons are more highly structured, as in English at Dalestone School, there is evidence that pupils can obtain better residuals and this is an area that would benefit from further research. Schagen

and Morrison (1999:8) concluded that the overachievement of higher ability students and underachievement of the less able *may* (their emphasis) be explained by pedagogical treatment of students in different sets. This is supported by the teacher interview data from this study. However, the scatterplots that they produce clearly demonstrate the Critchlow-Rogers Effect, and consequently the wisdom of inserting a caveat in their conclusion.

Where sets are shared, as in Science at Dalestone School, there is evidence that this has the effect of dampening residuals; that is, they tend to be closer to zero. This may indicate the effect of one less effective teacher bringing down the effect of another, more successful, colleague. However, it may demonstrate that these 'split classes' experience some other learning advantage or disadvantage when compared to those taught by a single teacher. This is an area worthy of further research.

The fourth area is that of teacher effectiveness. This research has shown that for statistical reasons it is not yet possible to calculate a number that reliably measures a teacher's effectiveness. There are also many other reasons why this is so. Interview data unveiled a number of factors both within and outside the control of teachers that can impact in positive and negative ways on the results that their classes obtain – but in teaching can you isolate a single factor?

In my discussions with teachers both in interviews and informal situations I have come to appreciate that there are also philosophical reasons why it is not desirable to reduce their work to a single number. Just as Ball (1997a:317) has

described schools as 'inherently paradoxical...complex, contradictory, sometimes incoherent organisations', so the same terms can often be used of teachers. Whilst many teachers now accept the current accountability agenda they do so having determined their 'room for manoeuvre'. Thus there is a tendency for them to be constructive for the purposes of evaluation and this process therefore to some extent invalidates the resulting measurements. Consequently, value-added measures at subject level are respected but at the level of individual classes teachers frequently, and often with justification, seek to influence the data.

Yet whilst we cannot come up with a useful number, we can see quite clearly who is doing well and who is not, at least in situations where there is fine setting or where there are parallel groups where we can be sure that the sets are genuinely equivalent. When we look at the average standardised residuals for each set, discontinuities in the expected pattern as predicted by the Critchlow-Rogers Effect show up stronger and weaker staff fairly clearly, particularly when the same observations appear year after year.

Unfortunately the above approach, whether setting is broad or fine, only gives information about relative effectiveness within the school. We still have no reliable way of producing a corrected average standardised residual that places the group and their teacher in a national context.

8.5 Implications for schools

This research has revealed several matters of importance in relation to the use and management of value-added data in schools. Schools have had to respond to a rapidly moving policy agenda. Having to account for 'raw' examination results only in the 1980s has now grown to a wide range of items including the Autumn Package and PANDAs. At Dalestone School one member of the Senior Management Team has managed the presentation of performance data in an authoritarian way that has had both negative and positive effects.

Many teachers have responded unenthusiastically to the imposition of detailed analyses and publication of performance data. Scepticism is not uncommon and, as has been observed with other initiatives such as the National curriculum, there is a period during which many teachers are 'reactive' to the change. However, at Dalestone School a significant number of those interviewed demonstrated a more positive, *heuristic*, attitude indicating both an acceptance of the new regime and a growth of professional confidence. This has been accelerated by the Threshold Assessment process that has encouraged reflection on practice together with a greater acceptance and use of data.

There is, however, too much data. The 'paralysis by numbers' syndrome results from the wide range of performance measures that exist. Some have national or LEA coverage whilst others are peculiar to individual schools. Where this data is used in the assessment of teacher effectiveness it can only be fair if

comparisons against national data are used. However, such data needs to be of good quality. 'Quick and dirty' (Thrupp 1999:155) approaches (such as the PANDA) which use aggregated data contextualised for socio-economic status, are of very limited use given the 'in-school' variation that has been highlighted in the present study.

Average residuals across setted groups may vary from year to year but generally the pattern reflects the policy by which pupils were set. There is a need to interpret residuals with care. First, they need to be understood technically, and this includes awareness of the Critchlow-Rogers Effect. Second, teachers bring their own interpretations and data are sensitive enough to show the effects of a variety of variables. Teachers identified issues that had an effect on value-added scores. These included pupil behaviour, changes to setting arrangements in response to budget cuts, and their own performance.

When data for individual teachers is examined over time it is possible to see that some teachers consistently obtain better results, for example, with lower sets. Whilst the data need to be treated with caution, they do have value in raising questions about performance and the deployment of staff. Low ability groups might obtain better residuals when they are allocated teachers who have a personal background of disadvantage or who have worked with disadvantaged youngsters.

8.6 Implications for further research

This research has pioneered an examination of the issues surrounding the use of value-added data in the assessment of teacher effectiveness. Although some reference has been made to the work of other schools, it has concentrated on one institution. Thus there is a need to investigate whether the findings relating to Dalestone School have a general application. This work has concentrated on core subjects where pupils are grouped by ability. There is scope for a broader appraisal of value-added data at class level in both setted and mixed ability groups across a wider range of subjects. In view of the current interest in Post-16 education, the policy and practice dimensions of value-added at this stage merit investigation.

During teacher interviews, several variables were identified that are worthy of further investigation. The issue of gender is a contemporary issue – just as some teachers seem to be more successful with pupils of a particular ability, are some teachers more successful with pupils of one gender or the other? Similarly, the effects of the timetable, group size, resources and other factors have all been subjects of research at some stage, but not specifically in terms of value-added outcomes. In the quest to raise achievement it is likely that future improvements will come about by smaller increments that result from a process of fine-tuning educational provision. Data about the variables that have been highlighted will be valuable in this process.

The present study has concentrated on relatively simple analyses of numerical data for reasons that were given in chapter 2. Larger scale research would be necessary to add a more rigorous statistical dimension to this work. Although this work has demonstrated the value-added bias that occurs in setted situations no correction method has been found. That is, we have no way in which teachers can be provided with data that accurately reflects the progress made by their teaching sets.

There is a need to pursue a way of correcting value-added data for setted groups which is free of the Critchlow-Rogers Effect. In addition to the use of national subject-level data in the calculation of residuals, other methods might be explored. First, an alternative, orthogonal, regression method, although mathematically complex, may more accurately reflect actual pupil progress compared to the national average. Second, a 'before and after' method could be used in English (where there is only one tier of examinations) or where data from tiered subjects could be adjusted to a single scale. Here, National Attainment Test scores can be plotted against YELLIS test scores at the start of Key Stage 4, and standardised residuals computed. The usual procedure is applied two years later with pupils' GCSE grades. If the average standardised residual of a teaching group increases during the two years the group can be said to have made better progress than the national average, and *vice versa*.

This study has found limited discernable influence of classroom practice on value-added policy. Yet this might be expected, given the stage in development that the school has reached. Accordingly, a follow-up study in two or three years

would be able to ask how much more are teachers able to exert their authority and find spaces for manoeuvre as was observed in studies of the National Curriculum. Can they make it work for themselves, that is, in their own interests, and in the interests of their pupils; and to what extent has the concept of value-added become internalised?

There are early signs at Dalestone school that some teachers are seeking ways of using value-added data in the quest to raise achievement. One idea is to use the YELLIS test data as a tool for early intervention. Thus, in Year 10 after YELLIS test results are received, pupils can be identified as belonging to the red, yellow or turquoise zones as described in chapter 5. As a result, those likely to obtain negative residuals (red zone) have a weakness in the subject in question but strengths elsewhere. They may need particular subject-specific support. Conversly, the turquoise zone pupils have a strength in the subject but weaknesses elsewhere and may benefit from a different kind of support.

This sort of development obviously needs encouragement and reflects a growing need for professional development in the use of value-added data. At one level there is a need for professional development in the management of data so that its potential is better realized. At the same time classroom teachers need more than to be told to make use of the numbers – their response is typically 'how?'

8.7 Teaching - Science or Art?

Reynolds' (1998b:26) notion that we should be 'pulling the "lever" of the teacher' has been under close scrutiny throughout this thesis. Value-added can be seen as an outcome measure of his science of teacher effectiveness. As such, it might represent the cold edge of accountability from a technology of teaching.

At a minimal level, there is surely a 'science' of teaching since there are things that a teacher can do to improve his or her general performance and these might only be recognised by the application of a scientific methodology. Beyond that, discussions about value-added have revealed a complexity and depth of emotion and passion amongst teachers. This 'artistic' view of teaching, which was described by Stenhouse (1985:106) as a 'personal construction', is immeasurable.

One of the most encouraging findings from this research is that many teachers are demonstrating a *heuristic* attitude towards performance data. In particular, they are engaging with it at the level of individual children and asking questions. The questions may have a scientific frame such as, 'am I differentially effective for a particular group of children?', followed by 'why?' The latter delves into areas that might include personal experiences, background, likes and dislikes, weaknesses or creativity. The 'mere number' then emerges from a frosty, constricted domain of accountability into the warmth and intricacy that are hallmarks of this profession.

Yet to debate whether teaching is a science or an art is really to miss what the profession is all about. It is about learning. What is important is that learning can be exciting, stimulating and creative. In order to engage with that process, teachers need tools to help them to focus on what is educationally important. One such tool is value-added.

Appendix

Semi-structured interview schedule

A note was sent to colleagues explaining the nature of the research and inviting them to be interviewed. An outline of the interview process and a note about confidentiality/anonymity was also given.

Before each interview commenced a reminder was given about the process and about how the data would be used.

Questions:

1. To start would you please outline your career to date - include any experience outside of teaching.
   Probes:
   Why did you decide to become a teacher?
   What do you find most attractive about teaching as a career.?

2. Value added data for key stage 4 has been collected at this school for a number of years. What is your knowledge and experience of value added?

   Probe - have you used value added data in planning your own lessons
   Do you feel that you in any way influence the way in which value-added data is used in the school?

3. In relation to *(a specific teaching set chosen by me - class list provided as an aid to memory)* tell me about how you taught them - the style you adopted, the resources used - and how you got on with them.

   Probes - Expectations, pace, pressure to succeed, level of work/tasks, differentiation, perceptions of individual pupils.
   Pay particular attention to any anecdotes.

4. Repeat question 3 for a different set.

5. What is your reaction to the Government's Green Paper proposing changes to the teaching profession - in particular the proposals about appraisal in which at least one objective should involve target-setting for expected pupil performance and/or progress?

   Probe - re green paper - what are your ideals and vision for the future?
   Green Paper 1:11 - 1:13. How could value added help teachers with 1:13 if at all?
   Probe - is it fair to relate a teacher's salary to pupil performance?

6. If the proposals contained in the Green Paper become law - particularly that relating to one element of your appraisal being linked to pupil progress - do you think that your teaching will benefit or will it be compromised?

   Probe - If we were to go back to the set discussed in question 3 what would you have done differently if an element of your appraisal was based on their progress?

7. What do you feel about the amount of autonomy or room for professional judgement that you have when inside your classroom?

   Probe - What do you think constrains your teaching?
   Does the National Curriculum act as a constraint?

8. In relation to the sets discussed earlier - how would you rate your performance?

9. In relation to the sets discussed earlier - how would you rate the performance of the pupils?

10. Repeat questions 8 and 9 for a different set.

11. Tell me about the ways in which you adjust your style in relation to different sets?

    Probes - In what ways do you change your style? Why do you change your style.
    Look for anecdotes which demonstrate changes of style linked to specific circumstances.

12. In relation to the teaching sets that we have discussed here are their value added results *(show data)*. What is your reaction to these figures?

    Probe - What reasons would you give for ...?..(pick out individual pupil or specific class results).

13. What do you understand by effective teaching?

    Probes - When do you do your most effective teaching?
    With what sort of pupil grouping do you feel most comfortable/effective?
    How do you know whether or not you are effective?

14. Finally, in this study I am trying to establish which teacher behaviours/attitudes/styles have most effect on pupil progress. Is there anything important about these aspects of your work that you feel that we have not had an opportunity to discuss and that you would like to talk about now?

School Records

Unpublished documents and other sources

Curriculum and Quality Development Team	1997	Key Indicators: support for School Self Improvement, Dalestone School	Chalkshire LEA
Dalestone School	1997	School Development Plan 1997-8	Dalestone School
Dalestone School	1998	School Development Plan 1998-9	Dalestone School
Dalestone School	1999	Responses to 'Targets 2000' document by Heads of Department. 17-09-99	Dalestone School
Dalestone School	2000	GCSE Subject Reviews – Subject Leader reflections on GCSE results.	Dalestone School
Dalestone School	2001	GCSE Subject Reviews – Subject Leader reflections on GCSE results.	Dalestone School
Deputy Head	1998a	Value-added measures. GCSE Results 1998	Dalestone School
Deputy Head	1998b	Targets 2000	Dalestone School
Deputy Head	1998c	Issues arising from the Target Setting INSET day. 01-11-98	Dalestone School
Deputy Head	1999	Key Stage 4 2000: Progress Check. May 1999	Dalestone School
Elsom, D	1995	Letter to Senior Teacher 12-10-95	
Former Headteacher	1995	Memorandum to Staff 05-01-95	Dalestone School
Governing Body	1999	Minutes from Spring Term meeting (January 1999) of the Full Governing Body	Dalestone School
Governors' Curriculum Committee	1998	Minutes from November 1998 meeting of the Curriculum Committee	Dalestone School
Heads of Department	1998	Public Examinations 1998. Subject Reviews by Heads of Department. 18-09-98	Dalestone School
Key Stage 5 Learning Co-ordinator	1998	Report on the CEM/YELLIS Conference. 03-11-98	Dalestone School
KS4 Reports	1999	KS4 Subject Leader/Heads of Faculty Reports. October 1999	Dalestone School
OFSTED/DfEE/QCA	2000	PANDA Report for Dalestone School	Ofsted/DfEE/QCA
Pupil MG	1999	Entries in pupil's record file	Dalestone School
Raising Achievement Working Party	1994	Raising Achievement Report. June 1994	Dalestone School
Rogers, S	1995	Notes from staff meeting led by David Elsom (CEM Centre) 10-01-95	Dalestone School
Rogers, S	1998	Notes from Target Setting Training Day. 23-10-98	Dalestone School

Rogers, S	1999	Notes from 'link' Department meetings. October 1999	Dalestone School
School A	2000	Data from Head of a West Midlands School. This school had been featured in a *Times Educational Supplement* article. March 2000	
School B	2000	Data from Head of a North Yorkshire School. Contact had been made via the CEM Centre Durham. 12-11-00	
Science Teacher	2000	Threshold Assessment Application. June 2000. Used with permission	Dalestone School
White, C	1996	Setting Targets for the Chalkshire Community: a shared commitment to raising attainment. 17-11-96	Chalkshire LEA
White, C	1997a	Notes for meeting of the Development Group: monitoring Pupils' progress at Key Stages 3 and 4. 22-01-97	Chalkshire LEA
White, C	1997b	Setting Targets to Raise Achievement. Subject Target-setting at Key Stages 3 and 4. 22-06-97	Chalkshire LEA
White, C	1997c	Standards of Attainment in Secondary Schools. Analysis of 1997 A-level and GCSE Examination Results	Chalkshire LEA
Woodward, A	1993	Value-Added Analysis - Performance at GCSE. 29-01-93	Rockside LEA

Interview Records

The sources listed here comprise interviews and conversations other than those that were formal semi-structured interviews with teachers.

Rogers, S	1998	Conversation with Head of Business Studies. 16-11-98	Dalestone School
Rogers, S	2001a	Conversation with Head of Science and Technology Faculty. 22-03-01	Dalestone School
Rogers, S	2001b	Regular personal communication and meetings with staff at the CEM Centre and with John Critchlow. 1999-2001	
Former Senior Teacher	24-06-00	Interview	
Former Headteacher	16-08-00	Interview	

References

Abraham, J. (1995). Divide and School: Gender and class dynamics in comprehensive education. London, Falmer Press.

Arksey, H. and Knight, P. (1999). Interviewing for Social Scientists. London, Sage.

Ball, S. (1981). Beachside Comprehensive: A Case-Study of Secondary Schooling. Cambridge, Cambridge University Press.

Ball, S. (1990). Management as a moral technology: a Luddite analysis. In S. Ball, ed. Foucault and Education: Disciplines and Knowledge. London, Routledge.

Ball, S. (1994). "Some reflections on policy theory: a brief response to Hatcher and Troyna." Journal of Education Policy 9 (2): 171-182.

Ball, S. (1997a). "Good School/Bad School: paradox and fabrication." British Journal of Sociology of Education 18 (3): 317 - 336.

Ball, S. (1997b). Markets, Equity and Values in Education. In R. Pring and G. Walford, eds. Affirming the Comprehensive Ideal. London, Falmer Press: 69-82.

Barber, M. (1996). "How to achieve the impossible." Times Educational Supplement (13th December, 1996): 19.

Barnard, N. (2000a). "Blueprint for the ideal teacher." Times Educational Supplement (23rd. June, 2000): 21.

Barnard, N. (2000b). "Secondary moderns 'do best of all'." Times Educational Supplement. (2nd. June, 2000) www.tes.co.uk

Bartholomew, H. (2001). Positioning students in setted mathematics groups. British Educational Research Association Annual Conference, Leeds University.

Bassey, M. (1999a). Case Study Research in Educational Settings. Buckingham, OUP.

Bassey, M. (1999b). "Performance-related pay for teachers: research is needed." Professional Development Today 2 (3): 15-28.

Baxendale, W. K. (1996). Inspection Report: Dalestone School, OFSTED: 37pp.

BERA (2001). "Report on Methodological Seminar on Hay/McBer Enquiry into Teacher Effectiveness: 9 May 2001." Research Intelligence (British Educational Research Association Newsletter) (76): 5-9.

Bernstein, B. (1970). "Education cannot compensate for society." New Society 15 (26th February, 1970): 344-347.

Black, P. and Wiliam, D. (1998). Inside the Black Box: Raising standards through classroom assessment. London, Kings College, London.

Blair, T. (1996). "Ruskin College Speech." Ruskin College. www.ruskin.ac.uk/archives/index.htm

Boaler, J. (1997a). Experiencing School Mathematics: teaching styles, sex and setting. Buckingham, Open University Press.

Boaler, J. (1997b). "Reclaiming School Mathematics: the girls fight back." Gender and Education 9 (3): 285-305.

Boaler, J. (1997c). "Setting, Social Class and Survival of the Quickest." British Educational Research Journal 23 (5): 575-595.

Boaler, J. (1997d). "When even the winners are losers: evaluating the experiences of 'top set' students." Journal of Curriculum Studies 29 (2): 165-182.

Boaler, J., Wiliam, D. and Brown, M. (2000). "Students' Experiences of Ability Grouping - disaffection, polarisation and the construction of failure." British Educational Research Journal 26 (5): 631-648.

Bourner, T., Bowden, R. and Laing, S. (2001). "Professional Doctorates in England." Studies in Higher Education 26 (1): 65-83.

Bowe, R., Ball, S. and Gold, R. (1992). Reforming Education and Changing Schools: case studies in policy sociology. London, Routledge.

Brown, M., Millett, A., Bibby, T., et al. (2000). "Turning Our Attention from the What to the How: the National Numeracy Strategy." British Educational Research Journal 26 (4): 457 - 471.

Burgess, R. G. (1983). Experiencing Comprehensive Education: A study of Bishop McGregor School. London, Methuen.

Burgess, R. G. (1984). In the Field: An Introduction to Field Research. London, Allen and Unwin.

Burgess, R. G. (1985). Field Methods in the Study of Education. Lewes, Falmer.

Busher, H. and Harris, A. (2000). Subject Leadership and School Improvement. London, Paul Chapman.

Callaghan, J. (1976). Prime Minister's Speech, Ruskin College, Oxford.

Cassidy, S. (1998). "Few cheers for value added." Times Educational Supplement School and College Performance Tables 1998 : 4.

Cassidy, S. (1999). "Added value put on back burner." Times Educational Supplement School and College League Tables 1999 : 4.

Chitty, C. and Lawn, M. (1995). "Introduction: Redefining the Teacher and the Curriculum." Educational Review 47 (2): 139-141.

Claydon, J. (1999). "Time to rethink and unite on pay." Times Educational Supplement (2nd April, 1999): 14.

Coe, R. and Fitz-Gibbon, C. T. (1998). "School Effectiveness Research: criticisms and recommendations." Oxford Review of Education 24 (4): 421-438.

Cohen, L. and Manion, L. (1994). Research Methods in Education. London, Routledge.

Coleman, J. S., Campbell, C. J., Hobson, C. J., et al. (1966). Equality of Educational Opportunity. Washington, US Government Printing Office.

Creemers, B. P. M. (1994). The Effective Classroom. London, Cassell.

Cullingford, C. (1995). The Effective Teacher. London, Cassell.

Curriculum Evaluation and Management Centre (1994). ALIS Newsletter. Newcastle, University of Newcastle upon Tyne.

Curriculum Evaluation and Management Centre (1999). The Yellis Handbook. Durham, CEM Centre, University of Durham.

Cutler, T. and Waine, B. (1999). "Rewarding Better Teachers?" Educational Management and Administration 27 (1): 55-70.

Davies, B. and West-Burnham, J. (1997). Reengineering and Total Quality in Schools: How to Reform and Restructure your School to Meet the Challenge of the Future. London, Pitman Publishing.

Day, C., Fernandez, A., Hague, T. E., et al. (2000). The Life and Work of Teachers: International Perspectives in Changing Times. London, Falmer.

Dean, C. (2000). "Grammars 'add less value'." Times Educational Supplement. (3rd. March, 2000) www.tes.co.uk

Dean, C. (2001). "Damning verdict on performance pay." Times Educational Supplement (13th July, 2001): 1.

Dean, C. and Thornton, K. (1999). "Children caught in north-south gap." Times Educational Supplement (10th December, 1999): 8.

Deem, R. (1994). "Free Marketeers or Good citizens? Educational Policy and Lay Participation in the Administration of Schools." British Journal of Educational Studies 42 (1): 23-37.

DFE (1992). Choice and Diversity - A new framework for schools. London, HMSO.

DFE (1995). Value Added in Education: A Briefing Paper from the Department for Education. London, Department for Education.

DfEE (1997). Excellence in Schools. London, HMSO.

DfEE (1998). teachers: meeting the challenge of change. London, The Stationery Office.

DfEE (1999a). Performance Management Framework for Teachers. London, Department for Education and Employment.

DfEE (1999b). Social Inclusion: Pupil Support. London, Department for Education and Employment.

DfEE (2000). Threshold Assessment application pack. London, DfEE.

DfEE (2001). Key Stage 3 National Strategy. Management Guide: Lessons from the pilot. London, Department for Education and Employment.

DfEE Standards and Effectiveness Unit (1997). From Targets to Action: Guidance to support effective target setting in schools. London, Department for Education and Employment.

DfEE and OFSTED (1996). Setting Targets to Raise Standards: a survey of good practice'. London, Department for Education and Employment.

DfEE Standards and Effectiveness Unit (1997). Setting Targets for Pupil Achievement: Guidance for Governors. London, Department for Education and Employment.

Donnelly, J. F. and Jenkins, E. W. (2001). Science Education: policy, professionalism and change. London, Paul Chapman.

Drew, R. (2000). Inspection Report: Dalestone School, OFSTED: 60pp.

Edwards, T. (1998). "Critique: Reynolds trivialises the complexity both of the means and the ends of effective learning." Research Intelligence (66): 29-30.

Elliott, J. (2000). "Revising the national curriculum: a comment on the Secretary of State's proposals." Journal of Education Policy 15 (2): 247-255.

Finkelstein, B. (1997). "Policy and practice in multiple perspective: case by case revelations in three nations and multiple sites, an introduction to this issue." Journal of Education Policy 12 (5): 309-311.

Fitz-Gibbon, C. (1992). School Effects at A -Level: Genesis of an Information System. In D. Reynolds and P. Cuttance, eds. School Effectiveness: Research, Policy and Practice. London, Cassell: 96-120.

Fitz-Gibbon, C. (1996). Monitoring Education: Indicators, Quality and Effectiveness. London, Cassell.

Fitz-Gibbon, C. (1997). "The Value-Added National Project: Final Report. Feasibility studies for a national system of value-added indicators." SCAA. www.qca.org.uk/education3-16/durham-report.htm

Fitz-Gibbon, C. (1998). "The value of value added." Managing Schools Today 8 (2): 23-24.

Fitz-Gibbon, C. and Kochan, S. (2000). School Effectiveness and Education Indicators. In C. Teddie and D. Reynolds, eds. The International Handbook of School Effectiveness Research. London, Falmer Press: 257-282.

Fitz-Gibbon, C., Tymms, P. B. and Hazelwood, R. D. (1989). Performance Indicators and Information Systems. In D. Reynolds, B. P. M. Creemers and T. Peters, eds. Proceedings of the First International Congress, London 1988. Cardiff, University of Wales and RION: 141-152.

Gewirtz, S., Ball, S. J. and Bowe, R. (1995). Markets, Choice and Equity in Education. Buckingham, Oxford University Press.

Gibson, A. and Asthana, S. (1998a). "School Performance, School Effectiveness and the 1997 White Paper." Oxford Review of Education 24 (2): 195-210.

Gibson, A. and Asthana, S. (1998b). "Schools, Pupils and Examination Results: contextualising school 'performance'." British Educational Research Journal 24 (3): 269-282.

Goldstein, H. (1983). "Measuring Changes in Educational Attainment over Time: Problems and Possibilities." Journal of Educational Assessment 20 (4): 369-377.

Goldstein, H. (1997a). "From raw to half baked." Times Educational Supplement. (18th July, 1997) www.tes.co.uk

Goldstein, H. (1997b). "Methods in School Effectiveness Research." School Effectiveness and School Improvement 8 (4): 369-395.

Goldstein, H. (1997c). "Value added tables: the less-than-holy grail." Managing Schools Today 6 (6): 18-19.

Goldstein, H. (1998). "A Response to Gibson and Asthana." Oxford Review of Education 24 (4): 521-523.

Goldstein, H. (2001). "Using Pupil Performance Data for Judging Schools and Teachers: scope and limitations." British Educational Research Journal 27 (4): 433-442.

Goldstein, H. and Cuttance, P. (1988). "A note on national assessment and school comparisons." Journal of Education Policy 3 (2): 197-202.

Goldstein, H. and Thomas, S. (1995). "School Effectiveness and 'Value-Added Analysis." Forum (for promoting 3-19 Comprehensive Education) 37 (2): 36-38.

Gray, J. (1994). "Relative Values." Times Educational Supplement. (16th December, 1994) www.tes.co.uk

Gray, J., Goldstein, H. and Jesson, D. (1996). "Changes and improvements in schools' effectiveness: trends over five years." Research Papers in Education 11 (1): 35-51.

Gray, J. and Hannon, V. (1986). "HMI interpretations of schools' examination results." Journal of Education Policy 1 (1): 23-33.

Gray, J., Jesson, D., Goldstein, H., et al. (1995). "A Multi-level Analysis of School Improvement: Changes in Schools' Performance over Time." School Effectiveness and School Improvement 6 (2): 97-114.

Gray, J., Jesson, D. and Jones, B. (1986). "The search for a fairer way of comparing schools' examination results." Research Papers in Education 1 (2): 91-119.

Gray, J., Jesson, D. and Sime, N. (1990). "Estimating differences in the Examination Performance of Secondary Schools in Six LEAs: a multi-level approach to school effectiveness." Oxford Review of Education 16 (2): 137-158.

Gubb, J. (1999). "Researchers do deserve Woodhead's support." Times Educational Supplement (9th April, 1999): 18.

Hackett, G. (1999a). "Inspectors link poor results and poverty." Times Educational Supplement (10th December, 1999): 1.

Hackett, G. (1999b). "Size does matter in the classroom." Times Educational Supplement (15th October, 1999): 8-9.

Hackett, G. (1999c). "Tables to lose progress measure." Times Educational Supplement (23rd April, 1999): 6.

Haigh, G. (1999). "Stats: the way to do it." Times Educational Supplement (10th September, 1999): 21.

Hale, R. (2000). "No basis for grammar slur." Times Educational Supplement. (10th. March 2000) www.tes.co.uk

Halpin, D., Moore, A., Edwards, G., et al. (1999). Maintaining, Reconstructing and Creating Tradition in Education. British Educational Research Association Annual Conference, University of Sussex, Brighton.

Hammersley, M. (1999). "Letter to the editor: Shades of grey in teacher skills debate." Times Educational Supplement (27th August, 1999): 15.

Hammersley, M. and Atkinson, P. (1993). Ethnography: Principles in Practice. London, Tavistock.

Hardy, J. (1998). "Turning the tables on teachers." The Guardian (12th December, 1998): 7.

Hargreaves, A. (1994). Changing Teachers, Changing Times: teachers' work and culture in the post-modern age. London, Cassell.

Hargreaves, A. (1996). "Revisiting Voice." Educational Researcher 25 (1): 12-19.

Harris, A. (1998). "Effective Teaching: a review of the literature." School Leadership and Management 18 (2): 169-183.

Harris, A. (2000). "What works in school improvement? Lessons from the field and future directions." Educational Research 42 (1): 1-11.

Harris, A., Bennett, N. and Preedy, M. (1997). Organizational effectiveness and improvement in education. Buckingham, OUP.

Harris, A., Jamieson, I. and Russ, J. (1995). "A Study of 'Effective' Departments in Secondary Schools." School Organisation 15 (3): 283-299.

Hatcher, R. and Troyna, B. (1994). "The 'Policy Cycle': A Ball by Ball Account." Journal of Education Policy 9 (2): 155-170.

Hay McBer (2000). "Raising Achievement in our Schools: Model of effective Teaching. An interim report from Hay McBer on the research findings." DfEE. www.dfee.gov.uk/teachingreforms

Helsby, G. (1995). "Teachers' Construction of Professionalism in England in the 1990s." Journal of Education for Teaching 21 (3): 317-332.

Helsby, G. (1999). Changing Teachers' Work. Buckingham, Open University Press.

Helsby, G. (2000). Multiple Truths and Contested Realities. In C. Day, A. Fernandez, T. E. Hauge and J. Moller, eds. The Life and Work of Teachers. London, Falmer Press.

Helsby, G. and McCulloch, G. (1997). Teachers and the National Curriculum. London, Cassell.

Hill, P. W. and Rowe, K. J. (1996). "Multilevel Modelling in School Effectiveness Research." School Effectiveness and School Improvement 7 (1): 1-34.

Hopkins, D., Ainscow, M. and West, M. (1994). School Improvement in an Era of Change. London, Cassell.

Huberman, M. (1993). The lives of teachers. London, Cassell.

Ireson, J. (1999). Innovative Grouping Practices in Secondary Schools. London, Department for Education and Employment.

Ireson, J. and Hallam, S. (1999). "Raising Standards: is ability grouping the answer?" Oxford Review of Education 25 (3): 343-358.

Jencks, C. S., Smith, M., Ackland, H., et al. (1973). Inequality: A reassessment of the effect of family and schooling in America. London, Allen Lane.

Jesson, D. (1995). Value-Added Aspects of Managing School Effectiveness. In J. Bell and B. T. Harrison, eds. Vision and Values in Managing Education: Successful Leadership Principles and Practice. London, Fulton: 232-249.

Jesson, D. and Gray, J. (1991). "Slants on Slopes: using multi-level models to investigate Differential School Effectiveness and its impact on Puplis' Examination results." School Effectiveness and School Improvement 2 (3): 230-247.

Kelley, P. (1999). "Do teachers make any difference?" Times Educational Supplement. (19th March, 1999) www.tes.co.uk

Kilyon, M., Fitz-Gibbon, C. and Defty, N. (1997). YELLIS and School Improvement. Durham, Curriculum, Evaluation and Management Centre.

Lawlor, S. (1999). "Can good teaching be measured? - No." Times Educational Supplement (13th August, 1999): 13.

Levi-Strauss, C. (1970). The Raw and the Cooked: Introduction to a science of mythology:1. (translated from the French by John and Doreen Weightman) London, Cape.

Maden, M. (2000). "New Labour's Bonapartist Tendency." Times Educational Supplement (25th February, 2000): 24.

Mayston, D. and Jesson, D. (1988). "Developing Models of Educational Accountability." Oxford Review of Education 14 (3): 321-339.

McCulloch, G. (1998). Failing the Ordinary Child? The theory and practice of working-class secondary education. Buckingham, Open University Press.

McCulloch, G. (2000). The Politics of the Secret Garden: Teachers and the School Curriculum in England and Wales. In C. Day, A. Fernandez, T. E. Hauge and J. Moller, eds. The Life and Work of Teachers. London, Falmer Press.

McCulloch, G., Helsby, G. and Knight, P. (2000). The Politics of Professionalism: Teachers and the Curriculum. London, Cassell.

McPherson, A. (1997). Measuring added value in schools. In A. Harris, N. Bennett and M. Preedy, eds. Organizational effectiveness and improvement in education. Buckingham, Open University Press: 184-190.

Morris, E. (1999). "Labour has the lighter touch." Guardian Education (9th March, 1999): 5.

Mortimore, P. (1998). The Vital Hours: Reflecting on Research on Schools and their Effects. In A. Hargreaves, A. Lieberman, M. Fullan and D. Hopkins, eds. International Handbook of Educational Change. Dordrecht, Netherlands, Kluwer.

Mortimore, P. (1999). "The voice of concern." Guardian Education (26th January, 1999): 3.

Murphy, R. (1996). "Drawing outrageous conclusions from national assessment results: where will it all end?" British Journal of Curriculum and Assessment 7 (2): 32-34.

Nuttall, D., Goldstein, H., Prosser, R., et al. (1989). "Differential School Effectiveness." International Journal of educational Research 13 (7): 769-76.

O'Donoghue, C., Thomas, S., Goldstein, H., et al. (1997). "1996 DfEE study of Value-Added for 16-18 year-olds in England." Institute of Education, London. www.ioe.ac.uk/hgoldstn/

OFSTED (1995). Guidance on the Inspection of secondary Schools. London, HMSO.

Osborn, M., Croll, P., Broadfoot, P., et al. (1997). Policy into Practice and Practice into Policy: Creative Mediation in the Primary Classroom. In G. Helsby and G. McCulloch, eds. Teachers and the National Curriculum. London, Cassell: 52-65.

Ozga, J. (2000). Policy Research in Educational Settings: contested terrain. Buckingham, OUP.

Patten, J. (1993). House of Commons Hansard (26th November, 1993): Column 696.

Pollard, A. (1985). Opportunities and Difficulties of a Teacher-Ethnographer: A Personal Account. In R. Burgess, ed. Field Methods in the Study of Education. Lewes, Falmer: 217-234.

Power, S. and Whitty, G. (1999). "New Labour's education policy: first, second or third way?" Journal of Education Policy 14 (5): 535-546.

Prescott, J. (1997). Afterword. In R. Pring and G. Walford, eds. Affirming the Comprehensive Ideal. London, Falmer Press: 197-203.

Qualifications and Assessment Authority (1998). Value Added Measures in School Performance Tables: consultation paper. Hayes, QCA Publications.

Rafferty, F. (1999). "CBI warns pay reform will fail." Times Educational Supplement (23rd April, 1999): 3.

Raudenbush, S. (1984). "Magnitude of Teacher Expectancy Effects on Pupil IQ as a Function of the Credibility of Expectancy Induction: A Synthesis of Findings From 18 Experiments." Journal of Educational Psychology 76 (1): 85-97.

Reynolds, D. (1997). "Now we must tackle social inequality not just assess it." Times Educational Supplement. (21st March, 1997) www.tes.co.uk

Reynolds, D. (1998a). "The school effectiveness mission has only just begun." Times Educational Supplement. (20th February, 1998) www.tes.co.uk

Reynolds, D. (1998b). "Teacher Effectiveness: Better Teachers, Better Schools." Research Intelligence (66): 26-29.

Reynolds, D. (1999a). "Can good teaching be measured? - Yes." Times Educational Supplement (13th August, 1999): 13.

Reynolds, D. (1999b). "It's the classroom stupid!" Times Educational Supplement (28th May, 1999): 13.

Reynolds, D. (1999c). School Effectiveness, School Improvement and Contemporary Educational Policies. In J. Demaine, ed. Education Policy and Contemporary Politics. Basingstoke, Macmillan: 65-81.

Reynolds, D., Hopkins, D. and Stoll, L. (1993). "Linking School Effectiveness Knowledge and School Improvement Practice: Towards a Synergy." School Effectiveness and School Improvement 4 (1): 37-58.

Rogers, S. (1997). The Modern Languages Sub-culture, Doctor of Education Part 1 degree assignment (unpublished). Division of Education. University of Sheffield.

Rosenshine, B. (1971). Teaching Behaviours and Student Achievement. Slough, NFER.

Rutter, M., Maughan, B., Mortimore, P., et al. (1979). Fifteen thousand hours: secondary schools and their effects on children. London, Open Books.

Sammons, P. (1999). "School effectiveness: coming of age in the 21st century." Education Journal September 1999 (37): 25-27.

Sammons, P., Thomas, S. and Mortimore, P. (1997). Forging Links: Effective Schools and Effective Departments. London, Paul Chapman.

Saunders, L. (1997). Value-added principles, practice and ethical considerations. In A. Harris, N. Bennett and M. Preedy, eds. Organisational effectiveness and improvement in education. Buckingham, Open University Press: 191-204.

Saunders, L. (1999). "A Brief History of Educational 'Value Added': How Did We Get To Where We Are?" School Effectiveness and School Improvement 10 (2): 233-256.

Saunders, L. (2000). "Understanding schools' use of 'value-added' data: the psychology and sociology of numbers." Research Papers in Education 15 (3): 241-258.

Saunders, L. and Rudd, P. (1999). Schools' use of 'value-added' data: a science in the service of an art? British Educational Research Association Annual Conference, University of Sussex, Brighton, NFER.

SCAA (1994). Value Added Performance Indicators in Schools. London, School Curriculum and Assessment Authority.

SCAA (1997). Value added indicators for schools: a consultative paper by the School Curriculum and Assessment Authority. London, School Curriculum and Assessment Authority.

Schagen, I. (2000). Statistics for School Managers. Westley, Suffolk, Courseware Publications.

Schagen, I. and Morrison, J. (1999). "A methodology for judging departmental performance within schools." Educational Research 41 (1): 3-10.

Scheerens, J. (1992). Effective Schooling: Research, Theory and Practice. London, Cassell.

Scott, D. (1996). "Education Policy: the secondary phase." Journal of Education Policy 11 (1): 133-140.

Simon, B. (1991). Education and the Social Order 1940-1990. London, Lawrence and Wishart.

Simons, H. (1996). "The paradox of case study." Cambridge Journal of education 26 (2): 225-40.

Slater, J. (1999). "Whitehall kills off classroom creativity." Times Educational Supplement (17th. March, 2000): 12.

Sparkes, R. A. (1999). "Value-Added - an uncertain measure." Scottish Educational Review 31 (1): 21-34.

Stake, R. E. (1998). Case Studies. In N. K. Denzin and S. L. Yvonna, eds. Strategies of Qualitative Inquiry. London, Sage: 86-109.

Stenhouse, L. (1985). Research as a basis for teaching. London, Heinemann.

Sylvester, R. (1998). "Teachers Threaten Strike action over Payment by Results." Daily Telegraph (20th. July, 1998).

Talbert, J. E. (1995). Boundaries of Teachers' Professional Communities in U.S. High Schools: Power and Precariousnss of the Subject Department. In L. S. Siskin and J. W. Little, eds. The subjects in question: departmental organisation and the high school. New York, Teachers College Press: 68-94.

Thomas, S. and Mortimore, P. (1996). "Comparison of value-added models for secondary school effectiveness." Research Papers in Education 11 (1): 5-33.

Thomas, S., Sammons, P., Mortimore, P., et al. (1997). "Differential Secondary School Effectiveness: comparing the performance of different pupil groups." British Educational Research Journal 23 (4): 451-469.

Thompson, C. (1999). The EdD and the Researching Professional: the importance of self-awareness of position in gaining the first and sustaining the second. British Educational Research Association Annual Conference, University of Sussex.

Thrupp, M. (1999). Schools making a difference. Let's be Realistic. Buckingham, Open University Press.

Times Educational Supplement (1998a). "Editorial: As good as we'll get." (4th December, 1988) www.tes.co.uk

Times Educational Supplement (1998b). "Editorial: Few cheers for value added." (4th. December, 1988) www.tes.co.uk

TTA (1998). National Standards for Qualified teacher Status. London, Teacher Training Agency.

Tymms, P. (1993). "Accountability - can it be fair?" Oxford Review of Education 19 : 291-299.

Tymms, P. and Fitz-Gibbon, C. (1991). "A comparison of Examination Boards: A-levels." Oxford Review of Education 17 (1): 17-31.

Walford, G. (1997). Privatization and Selection. In R. Pring and G. Walford, eds. Affirming the Comprehensive Ideal. London, Falmer Press: 54-65.

Weick, K. (1976). "Educational Organisations as Loosely-Coupled Systems." Administrative Science Quarterly 21 (1): 1-19.

Wellington, J. (1996). Methods and Issues in Educational Research. Sheffield, University of Sheffield, Division of Education.

Wiliam, D. and Bartholomew, H. (2001). The influence of ability grouping practices on student achievement in mathematics. British Educational Research Association Annual Conference, Leeds University.

Willetts, D. (1999). "Time to set schools free." Guardian Education (9th March, 1999): 4.

Williams, J. and Ryan, J. (2000). "National Testing and the Improvement of Classroom Teaching: can they coexist?" British Educational Research Journal 26 (1): 49-73.

Woodhead, C. (1998). "Subliminal messages." Times Educational Supplement (20th November, 1998): 13.

Woods, P. (1996). Researching the Art of Teaching. London, Routledge.

Yang, M., Goldstein, H., Rath, T., et al. (1999). "The Use of Assessment Data for School Improvement Purposes." Oxford Review of Education 25 (4): 469-483.