Chapter 2 Aims and Scope

Primary Aims

I began this research following a request from the newly appointed Head of Sexey's School that I establish how well the school was doing in terms of examination results. This apparently simple request hid a multiplicity of problems involved in the establishment of school effectiveness, many of which I will discuss in a review of the current literature in the next chapter of this thesis.

The need to establish Sexey's School's effectiveness resulted from many of the problems outlined in the introduction to this thesis. The forthcoming publication of Secondary School Performance tables (DFE, 1992) placed immediate emphasis on the need for the school to "do well". Linked to this was the need to gain a better understanding of our pupil intake and what they should be capable of achieving.

As a new Head to the school it was important for him to be able to show that the examination results were improving or, at the very least, not declining since he took up the post. This was increasingly important as Governors were given greater control over the running of the school by government legislation, including the right of hiring and firing staff, and the likelihood of Headteachers having to negotiate their own salaries based upon their performance which inevitably meant the performance of the school.
Grant Maintained Status was on the horizon and pupil recruitment in a small school was always potentially critical to the school's survival. Worries about a bad showing in the "League Tables" and possible negative effects upon recruitment were to the fore in the minds of many of the staff.

• My first aim was to explore ways of establishing what the examination performance of the school was. As Examinations Officer at the time I already had to hand various statistics on examination results such as overall pass rates

(percentage of pupils achieving grades A-E for Advanced level and A-C for GCSE level) as well as figures for percentage of candidates achieving various grades and so on.

These data whilst reflecting the achievement of each year cohort gave very little information about the performance of the pupils, in relation to what might be expected of them, or the subject teachers or indeed the school as a whole.

In order to tell whether the school was maximizing the academic potential of its pupils or the reverse it would be necessary to establish some baseline against which to judge performance in examinations. For A level results it seemed logical to look at pupils' performance in GCSE subjects for some indication of likely expectations at A level. GCSE results being summative assessments of the pupils' educational progress thus far and also a minimum standard had to be achieved in these examinations for admission to the Sixth form. The Sixth form at Sexey's School comprises some 200 pupils of whom at least half are new to Sexey's when they arrive in the Sixth form. As I was seeking a measure of prior attainment for all these pupils the GCSE results represented the only indicator common to all pupils.

For GCSE performance I chose to use the pupils' scores in the Edinburgh Reading Test (ERT) Stage 4 which all pupils sat in the Spring term of Year 8. The test was standardised for age with a mean score of 100. This would hopefully enable me to compare pupils of different ages and, if necessary pupils who sat the test at different times.

If I could demonstrate a relationship between my indicators (GCSE results / ERT) and outcomes (A level results / GCSE results) then I could demonstrate what is now commonly known as "value-added", the difference between what might have been expected from pupils with said scores and what was actually achieved.

However, in order to achieve my first aim it was important to establish the efficacy of the links between indicators and outcomes, the degree to which outcomes were related to indicators, the extent to which any relationship found could be generalised to other pupils, year groups or schools.
It was not uncommon for A level subject teachers to consider their pupils' GCSE result in their particular subject and find very little link if any between the GCSE result and the pupil's eventual A level result. Similarly, many teachers at Sexey's, myself included, were dubious about what a reading test could tell us about pupils' potential in French, Maths or Science GCSE examinations.

• In consequence of these concerns, my second aim was to explore the strength of any relationship between my indicators and outcomes. If the strength of the relationships were significant then I would be able to draw some positive conclusions about performance in relation to expectations. I would hope to be able to convince my colleagues of the utility of my work. Establishing the predictive validity of the indicators became a priority. This could be measured by calculating the co-efficients of correlation between indicators and outcomes. The higher the correlation co-efficient the stronger the relationship. Taking this a stage further it would be possible to calculate a co-efficient of determination by squaring the correlation co-efficient and arriving at a figure that would indicate the percentage of variance in the outcome that could be explained by the variance in the indicator.

There would also be an issue of reliability. It was important that the ERT prove reliable in distinguishing between pupils so showing the dependability of the test. It was also important that the relationship between indicator and outcome prove reliable. Consistency of strength in the correlations between indicator and outcome would be important in persuading staff of the utility of the process and yet to over-emphasise the importance of high correlation would be to deny the teacher / pupil credit for gaining high outcomes despite a low

indicator score.

Whilst high correlation was important for establishing the general relationship between indicator and outcome it does not of itself indicate whether one would interpret the set of results as being good or bad. In theory it would be possible for a set of examination results to be very highly correlated with the pupil indicator scores and yet the results to be poor in relation to what might be expected of those same pupils with a different teacher, in another school, in a different year. I would need to consider the interpretation of correlation figures very carefully.

• My third aim was to explore the ***implications*** of considering examination performance in relation to a baseline indicator. In relation to the school, my intention was to seek explanations for year on year variation in examination results for successive cohorts of pupils. I wished to determine whether supposed "good years" really were years of high performance in relation to expectations or had the staff expectations been falsely low. Indeed, were the pupils being stretched or being let down by the targets set for them by staff.

In the various subject areas I wished to consider departmental performance. In the traditional first staff meeting of the year, following the publication of the examination results, praise was given to departments whose pupils had gained high grades. The performance indicator which was being used to distinguish departments which had done well was the percentage of "A" grades achieved. This militated against departments with weaker intakes whose students could not be expected to achieve such grades. In order to determine whether a department had students who were weak or students who were strong academically, the establishment of some baseline for comparison of relative ability was essential. This would hopefully also allow for some comparison of subjects in terms of difficulty but considering only Sexey's pupils would not provide sufficient sample size for any general conclusions.

I wanted to explore the strength of relationship between indicator and examination outcome at the subject level and whether the indicator held good for all subject areas or just some, mindful that strength of relationship does not necessarily imply quality of results. I was also seeking some objective evidence regarding department and teaching staff performance rather than the general staffroom gossip which had pertained up to this point. If the relationship between indicators and outcomes remained sufficiently strong for individual subject areas then perhaps the teacher of the low ability groups and the pupils in those groups could at last gain some praise for their achievement rather than just those teachers who taught or were in the most able groups. Academic achievement could be judged in relation to expectations and praise given accordingly.

The reactions of the teaching staff to the indicator data and my findings would be important. I needed to be able to demonstrate the link between indicators and outcomes in a manner which they could accept. The presentation of any findings would need to be handled sensitively or the process could become threatening, which was not my intention, and could result in teachers fearing the data or even being hostile to it. I hoped to be able to provide objective evidence of real achievement by pupils and staff which could then be fed back into the educational processes operating on successive year groups coming through the school. At the very least open and honest discussion could be entered into over factors contributing to success and failure.

• My fourth aim was to explore the data for evidence of gender effects. I wished to ascertain if girls were performing at a higher level when the respective genders were compared with pupils of similar ability.
If there were any differences in performance overall could these be traced to particular subject areas or combinations of subjects?
In Sexey's School, because of the makeup of the boarding provision, there is always a gender imbalance in terms of pupil numbers in favour of boys. There are more beds for boys than girls. Given that girls are therefore outnumbered

by boys would they perform worse, as well as or better than the boys?
If either gender were found to perform significantly better than the other then this would highlight an important point in the understanding of the overall performance of the school. For instance, if girls outperform boys at GCSE then the relative numbers of girls in the year cohort could have a major impact on the headline figure reported in the National Performance Tables, that is the number of pupils (regardless of gender) achieving five or more GCSEs at grade 'C' or above.

• My fifth aim was to create a system that would allow for rapid and valid data analysis. The system would also have to provide teaching staff with comprehensible information that they perceived as being useful and would therefore use. If analysis of data took too long then the cost of providing time for a member of staff to do the work would militate against any system being implemented nor would any member of staff choose to do the necessary work in their own time. There was also the point that if examination results were to be reviewed with individual members of staff and the staff as a whole then the most opportune time is immediately at the start of the term following the release of those results when individual pupils and circumstances are fresh in the teachers' minds. A system that took months to produce any useful information would mean that this opportune timeslot would be missed.

Further Aims

As my work got under way a further set of aims gradually developed; an important one of which was to involve other schools' and their examination data. Involving other schools would provide data for comparison rather than restricting the research purely to one school. Each school would be able to compare its own data with that of the combined sample of other schools.
For reasons of confidentiality I decided to make available to all schools as part of the feedback the combined data rather than the data of each individual school. It would have been far too easy to ascertain to which school each data set belonged had I not done this and in the competitive atmosphere prevailing

amongst schools at the time many schools would not have released their data to me otherwise.

With more schools involved the size of the data set would increase and I would be able to check any findings I made at Sexey's against the wider sample. Also the larger the sample size the more sure I could be of the significance of any correlations. The larger sample would also allow me to consider the data at the level of individual subject areas rather than at the whole school level. As Sexey's is such a small school with a year 11 cohort size of around 50 each year, even at the whole school level, I would need to combine a number of years' results to achieve a data set of say 200 pupils.
At Sixth form level the Sexey's year 13 cohort is usually around 95 which is a better sample size than the year 11. However, the advantage of a sample size increased by incorporating the data from other schools meant that I could look at the data for subject areas where the pupil numbers within Sexey's School were small. Such a subject was German with only 7 pupils over the period 1994-1996 in Sexey's but a combined schools' sample size of 70 in 1996 and 145 pupils over the period 1994-1996.

Linked to the increased size of the data sets and possible greater statistical significance of any findings was the hope that the involvement of other schools would improve the Sexey's School teacher perception of the system I was developing. I hoped to be able to demonstrate that findings beginning to emerge at Sexey's were not unique but could be replicated in other schools and in larger pupil populations.
For those teachers with less of a statistical bent, the fact that other schools and teachers were prepared to become involved and by their involvement were seen to be voluntarily supporting the research, rather than having it inflicted upon them, was a very positive thing. It was also encouraging to me.

Other schools becoming involved, voluntarily, and being prepared to share information, all be it anonymously to everyone except myself, for the mutual

benefit of all schools involved was to be uplifting in a time of intense and occasionally acrimonious public competition. Here would be an example of schools "networking" - sharing information on performance in public examinations for their mutual benefit - some six years before the concept was officially promoted by the DFEE.

The inclusion of more schools in the research data set allowed me to pursue another aim, that of comparing the effectiveness of different schools in terms of the results achieved by their examination year cohorts.
The fact that the majority of schools submitting their GCSE results for analysis also used the Edinburgh Reading Test with their Year 8 pupils, and so had a common baseline indicator, meant that I could consider the value-added performance of each school in relation to the expected performance. The expected performance would be based upon the results of the combined sample of all schools with the same indicator information.
Not only did I hope to be able to consider comparison of whole-school performance but the performance of individual subject areas between schools. I would be able to compare English Language results in one school with English Language results in all schools, as much as possible a comparison of like with like. The range of variables to be taken account of in trying to compare like with like is large. Many of the problems involved in making comparisons of effectiveness are aired in the next chapter of this thesis where my research was informed by a review of the academic literature.

In considering the effectiveness of individual schools, and for that matter individual subject departments with different schools, using the pupil indicator information I hoped to be able to explore the effects of the distribution of pupil ability in the schools and departments. I was concerned that the average (mean) ability for the groups or departments might well hide further information that ought to be taken into account when considering school or departmental effectiveness. If the pupils of differing ability were not normally distributed in the school or department populations then the outcome measures

used to judge effectiveness would have to be chosen carefully. Statistical hurdles, such as the percentage of pupils achieving five or more GCSE grades at the level of A* to C grade could be very misleading if a large number of pupils were not sufficiently able to clear that hurdle. The same logic applies within a single subject area when considering the percentage of pupils attaining grade C or above. Comparison of two schools, or subject areas, with similar average abilities may well find that one has more pupils of an ability level capable of clearing the hurdle but less very able pupils. The other school or department might have a few very able pupils, so raising the group's average ability score, but many more less able pupils not realistically capable of achieving the required threshold grades or grade.

This issue leads on in turn to the question of critical levels of ability for success in examinations, however one chooses to measure success.
By looking at larger numbers of pupils with the same indicator data I hoped to be able to ascertain the critical level of ability which seemed to be required to achieve the various grades in examinations. Would there be consistency year on year in the same school or across different schools? If this could be established then the composition of school year cohorts and the distribution of pupil ability in them might be seen as being critical to achieving more traditional measures of success such as percentage of pupils gaining the various grades in GCSE and GCE A level examinations.
Furthermore, would critical levels of ability prove to be different for girls and boys with consequent implications for research into school effectiveness needing to look at the gender composition of year cohorts and the respective distribution of ability within the gender groupings.
These particular gender issues are illuminated later in this thesis in the case study of School X.

My final aspiration was to be able to link the findings from my analysis of examination data to current practice within Sexey's School for the benefit of those pupils currently coming through the school and future pupils.

Analysis of the data would hopefully be able to provide some indication of likely outcome for pupils of any given ability so that current academic performance could be considered in relation to likely outcomes, assuming roughly average performance, and targets set accordingly. Weaknesses and strengths in the teaching systems, such as pupil grouping arrangements, provision of syllabuses suitable for the particular pupils, allocation of staff to particular pupil groups and so on, would be highlighted in a more illuminative and accurate way than had been current practice so allowing appropriate action to be taken. The effectiveness of systems could be related to actual academic progress, taking account of the prior ability of pupils, rather than pupil attainment in isolation from other contributing factors, the primary one being prior ability.

Use of the information on year cohort ability and the prior test information would help when decisions were being made regarding the curricular provision required to best suit pupil needs. Individual pupil data could be used to counsel pupils regarding their subject options, their current academic performance in relation to expectations based on the performance of previous pupils with similar indicator information.

This use of prior test data must not, of course, ignore the very many other variables, including personal circumstances, which operate upon pupils and their chances of academic success. My study of the research literature would inform my understanding of these various factors and the extent to which they played a part in the eventual attainment of pupils and consequential performance of year cohorts and schools. The analysis of examination data is to provide information which will help the educational process and offer a diagnostic element complementary to the teachers' knowledge of individual pupils and their needs.

To summarize my last aim, I wanted to make formative use of summative assessments for the benefit of future examination candidates and their

teachers. Teachers would have a key role to play because of their unique knowledge of their pupils and factors beyond prior test scores.

Scope of the research data

My initial exploration of Sexey's School examination results began in 1990 looking at both GCSE results and A level results. Edinburgh Reading Test (Stage 4) scores were also available and I was interested in looking at the levels achieved by different groups and possible prediction variables.
I looked at previous years' results but the information was limited. No records had been kept of A level candidates' GCSE results prior to 1990 so this was the first year that comparisons could be made.
Edinburgh Reading Test (ERT) information for 'O' level candidates was available as far back as 1987 and so exploration began with the results for that year.

At this stage only basic statistical analyses were made on relatively small groups but the writer had to perform these manually which was extremely time consuming and prone to inaccuracies. It was very quickly apparent that more data could be analysed more quickly and with greater accuracy by buying or developing appropriate computer software. I decided to write my own software to answer the sorts of basic questions I was asking. One advantage of writing this software myself was that I could amend it at any time to suit my purposes better. Commercial software to do the same task was not available and spreadsheets struggled to handle the amount of data.

Although indicator data were available for 1987 'O' level results only 25 pupils out of a year cohort of 46 had ERT data. In successive years the proportion of candidates who had prior data improved and following 1991 steps were taken to ensure that the datasets were as complete as possible.

In 1991 six South Somerset secondary schools were persuaded to submit their

GCSE results for analysis. I needed the extra data from more schools to increase my dataset and further my research aims as already stated. The schools were relatively local, all within the South Somerset area, and therefore were aware of the geographical, social and financial factors operating within that area. Whilst each school could justifiably claim to be different from the other schools there were commonalities and in particular the common baseline data of Edinburgh Reading Test scores taken by the pupils at the same stage in their education.

Probably the most pressing reasons for schools joining in were the educational climate prevalent at the time, one of intense competition and market forces, and an element of mutual benefit. These two reasons may seem to be mutually exclusive but the anonymity of the research and data being fed back meant that market positions were not going to be threatened and if there was to be some advantage gained they all wanted to be party to it.

Each school sent either their Headteacher or Deputy to discuss the analysis of their own school's data. These discussions took place on a one to one basis with myself and usually lasted between an hour and one and a half hours. The Heads' and Deputies' initial scepticism of the process, the meaning of the analyses, concerns over the confidentiality of the material, trepidation about statistics were gradually dissipated. The relevance and utility of the information, particularly as it concerned individuals whom they knew, both staff and pupils, helped greatly.

In describing some of the casework developed within Sexey's I was able to convince them of its relevance to formative improvement. For instance, I was able to point to individuals who had apparently done badly in the examinations but had achieved a great deal in relation to their ability and how this could be related to the motivation, self-image and success of teachers who might otherwise have been discouraged. In other cases the scenario was the reverse and there were messages they quickly saw for school improvement.

The important point to make here is that the Head or Deputy who knew the individuals and their particular circumstances was able to confirm my judgements based on the value-added data. Gradually the potential of the information became apparent to them and enthusiasm took over, particularly where the statistics illuminated circumstances which the schools' representatives had felt to be the case but lacked the evidence to prove, or where their judgement had been seriously in error because no account had been taken of the ability of the pupils.

With some individuals this enthusiasm had to be tempered with warnings from myself about reading too much into the figures, statistical significance and correlation not implying causation. Headteachers and Deputies come from many different disciplines and many do not have a statistical background. They therefore approach statistics with a degree of temerity or are too ready to jump to conclusions without understanding the limitations of the statistics themselves.

Two schools immediately submitted previous years' results to build up a data set for their own schools. In 1992 nine South Somerset schools submitted their GCSE examination results following word of mouth recommendations. This gave a sample size of some 1135 pupils with matched pair data, GCSE results and ERT scores, a considerably better sample size than the 26 pupils at Sexey's that year.

Following an article by Gerald Haigh about the beginnings of this scheme published in the Times Educational Supplement (Haigh, 1992) a number of schools from outside Somerset also sent me their examination results and prior test scores for value-added analysis. These schools tended to use different prior tests and so their results were kept separate from the other schools as the baseline indicators were different.

In 1993 I was approached by Somerset LEA who were interested in working

on value-added analysis of GCSE results and had started to produce some data of their own. The LEA decided to take up a licence on my software and provide the analyses for all secondary schools in Somerset.
Despite having the data provided by the LEA most of the South Somerset schools chose to continue sending me their examination data directly because of the speed with which I was able to get the value-added information back to them and the possibility of having further specific analysis work done on issues such as gender.

Further publicity, Haigh (1994) and Tytler (1995), has seen the number of schools involved grow until in 1996 there were 27 schools submitting GCSE results of which 18 used ERT as their baseline. These 18 schools together give a sample size of 2834 pupils.
At A level the number of schools involved is less. In 1996 there were 12 schools giving an A level pupil sample of 1033.

Since the start of my work I have addressed a number of meetings on value-added analysis of examination results including the Grant Maintained Schools' Deputies Conference, Leicester, 1992; Boarding Schools' Association Deputies' National Conference, Webbington, 1992; Centre for the Study of Comprehensive Schools South-West conference, Bristol, 1993; Avon TVEI School Improvement Conference for Heads, 1994; Somerset Deputy Head's Conference, Dillington, 1995; Canford Group of Independent schools' Deputies' meeting, Truro, 1995; Department for Education, OFSTED and HMI, London, 1995.

In addition to the above presentations, I have delivered In Service Training to staff at a number of schools and also to a governing body. These visits in particular have been helpful to me in appreciating the problems faced by teaching staff and management teams in coming to terms with and understanding statistical data on examination results, interpreting them as measures of school performance and using the messages they give to help

focus and develop methods for school improvement.

My own understanding has developed in studying the widespread recent developments that have taken place in the areas of school effectiveness and school improvement. The use of different statistical methods for data analysis and educational arguments about their interpretation, use and educational value have now established a formidable amount of research literature. This is the subject of review in the next chapter.