The Value Added National Project

FINAL REPORT

Feasibility studies for a national system of Value Added indicators

THE VALUE ADDED NATIONAL PROJECT

FINAL REPORT

Feasibility studies for a national system of value-added indicators

Carol Taylor Fitz-Gibbon

February 1997

SHORT TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES

TABLES

EXECUTIVE SUMMARY

The objective of the Value Added National Project

The stated objective of the Value Added National Project was

> *"To advise the Secretary of State on the development of a national system of value-added reporting for schools based on prior attainment, which will be statistically valid and readily understood."*

This is the final report from the 22 month project.

Project methods

The work undertaken included

- statistical trialling;
- experiments on methods of providing data;
- pilot value-added reporting;
- surveys / questionnaires
- meetings with
  - teachers' associations;
  - schools participating in the pilot studies;
  - statisticians;
  - Local Education Authorities;
  - a value-added advisory group

The contract required a focus on Key Stage 2 and Key Stage 4.

Criteria and uses for a national value-added system

Criteria for a national value-added system were agreed as

- readily understandable;
- statistically valid;
- not an undue burden on schools;
- cost effective.

Two major applications of the system were considered:

- the use of value-added information for internal school management;
- the use for public accountability.

Definition of value-added

A simple definition of value-added provided a readily understandable approach. 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. Thus value-added for a key stage was essentially a measure of the pupil's relative progress during that key stage.

Since the average of prior attainment measures produced the best basis for prediction of each subject, a single figure kept on record for each pupil would provide the basis for value-added calculations in all subjects. This indicates a minimal requirement for record keeping to enable value-added measures to be computed.

Statistical analyses and survey databases

Despite some difficulties in locating schools that had retained the necessary end-of-Key Stage test results, datasets were analysed from more than 26,000 primary pupils from 753 school cohorts and more than 31,000 secondary pupils from 224 school cohorts (Table 1.2). In addition headteachers from 104 primary schools and 389 secondary schools responded to value-added reporting.

Statistical Findings

Average attainment levels from key stage tests provided an adequate basis for the calculation of value-added scores. Using the readily understandable definition of value-added scores as relative progress, and applying simple statistical methods, value-added indicators for schools showed no differences of practical significance from indicators calculated by more sophisticated and less accessible methods. The readily understandable methods were defensible and recommended for 'Stage 1' data analysis and feedback to schools. However, there were reasons to recommend a 'Stage 2' analysis prior to using the data for public accountability:

- some cohorts would be too small to provide reliable indicators for a single year's data and would need to be excluded prior to any use of the data for purposes of accountability;
- some schools with atypical intakes would be more fairly represented if additional factors (e.g. sex, English as a second language) were taken into account, possibly providing adjustments to the value-added scores;
- the newly available national datasets should be fully explored and analysed by several contractors;
- the impact of additional variables should be explored, particularly with knowledge of local conditions, to promote insights into school effectiveness.

Views of headteachers

The views of headteachers were very largely positive towards the use of value-added data within schools; it was not likely to be seen as a **burden** but as needed and useful, particularly if it was kept simple and understandable and used cautiously. There were grave reservations concerning publication although a majority of the secondary heads surveyed would want value-added indicators included in any School Performance Tables. A need was seen for INSET and other forms of help, for headteachers, governors and parents.

Organisational features of a national system

The **costs** of a national system will depend upon the extent to which electronic data interchange is used and unique pupil identifiers are available. Unique identifiers are needed to facilitate keeping track of all pupils and matching data from different schools, tests, examinations and years. If this necessary technical infrastructure is developed, the costs and timeliness can be most favourable.

Costs could be kept low, and feedback provided promptly, by a system of registration in which prior attainment data for each pupil accompanied registration of pupils for end of key stage tests or examinations.
If two sets of data, namely registration data (including prior attainment) and results data (including levels or grades and mark) were routinely copied to a central agency or agencies, then a national database could be constructed and made available for further analyses, focused particularly on information useful for school improvement. Furthermore these datasets could provide a framework for statistical monitoring of comparability between subjects, syllabuses and examination boards.
As the test and examination system is used to provide 'high stakes' information, it is important that such monitoring is thorough and there are improvements in quality assurance procedures in the test and examination process.

Effectiveness of a national value-added system

The effectiveness of a national value-added system will be dependent upon confidence in the integrity of the data underpinning the system and the careful interpretation of the data.
For internal school use, methods of monitoring and tracking pupils can be readily adopted by schools provided with readily understandable pupil-by-pupil lists of value-added scores or the simple provision of the national trend lines ('regression lines') for each syllabus or subject that has been assessed.
The use of value-added data for public accountability will require additional safeguards and, since value-added indicators vary substantially from year to year, could advisedly be postponed until three years' data become available. A major problem will be the need to take account of rates of pupil attendance and mobility. Various methods for the development of uniform, reliable systems were suggested including the possibility of some amount of discretionary non-reporting, subject to inspection to ensure reasonable application of criteria for non-reporting.
A value-added system suitable for pupils with special needs will need special research and development.
Earlier national publication could be considered if data were published only from the top 10 percent in each subject, region or examination. The national value-added system would thereby be used to give recognition to excellent progress and to provide benchmarking information that might inform practice in other schools.
This report makes 36 recommendations relating to general features of an initial system, the required infrastructure, quality assurance procedures for assessment methods, primary and secondary school value-added systems, and monitoring the implementation of the value-added system. Recommendations are in Section 7.

SECTION 1: INTRODUCTION

1.1 The Contract

The Value Added National Project resulted from a contract that stated one objective:

> *"To advise the Secretary of State on the development of a national system of value-added reporting for schools based on prior attainment, which will be statistically valid and readily understood."*

This is the final report of the 22 month project (March 1995 to December 1996). The first phase of the project, consisting of statistical analyses of existing datasets, was completed with two technical reports and an interim general report, all in December 1995. The second phase has resulted in five additional technical reports and this final report for a general audience. All nine reports are listed in Annex A.

The background to this final report is described in this introduction, which then concludes with an outline of the structure of the remainder of the report.

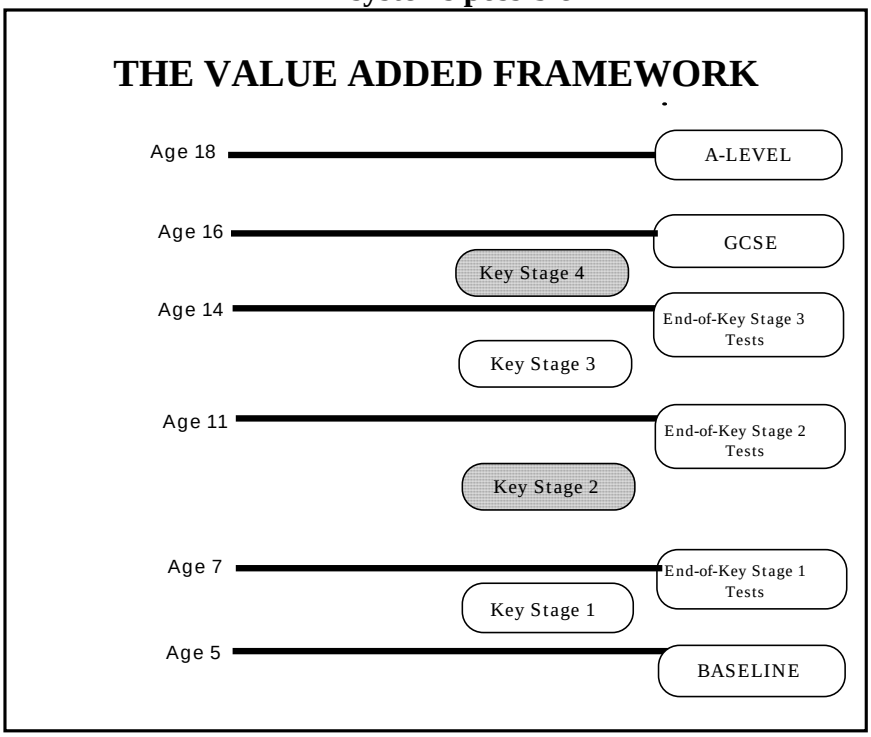
1.2 Background

This project was concerned with 'value-added' indicators relating to externally measured attainments. There are, of course, many ways in which a school 'adds value' in addition to the progress made by pupils in examination results. For example: pupils' social outcomes, skills in the arts, sports and vocational areas, life skills, interests, attitudes and aspirations. Schools also 'add value' to the lives of parents and the community.

> *At the outset, we must recognise that there are many ways in which schools "add value" to their pupils ..... the others are no less essential and must feature as part of the context of the whole school within which any value added measure must be interpreted.*
>
> Sir Ron Dearing in his Foreword to the first report

Externally awarded credentials, such as examination results or the "Key Stage" test results, are nevertheless a major responsibility of schools and a matter of concern to parents, employers, further and higher education. Value Added information will thus be important to all these audiences. With the introduction of the National Curriculum in English schools and a system of testing at the end of each key stage, the UK now has in place a framework of external assessment that provides the potential for Value Added measures to be produced at the end of every Key Stage, as indicated in Figure 1.1. The Value Added National Project was to look specifically at the end of Key Stage 2 and the end of Key Stage 4 (GCSE examinations).

Figure 1.1 The framework of tests and examinations that makes value-added systems possible



The Value Added Advisory Group, the membership of which is listed in Annex B, included representatives from schools, SCAA[1], DfEE, OFSTED and the CEM Centre, University of Durham. Members of the advisory group agreed that there were two major purposes for a value-added national system. One purpose was to provide important information for public accountability, additional to that provided by the School Performance Tables. A second purpose was to provide data for internal school management and to provide teachers, each year, with indicators of the progress of their pupils compared with the progress of similar pupils in other schools. Because of the usefulness of data provided pupil by pupil and subject by subject, thousands of UK schools are already participating in value-added systems.

The Value Added Advisory Group also agreed four criteria for a national value-added system. It should be:

- readily understandable;
- statistically valid;
- not an undue burden on schools;
- cost effective.

The first two criteria could well have been in conflict. However, extensive statistical trialling demonstrated that very simple and accessible models of relative progress (value-added) yielded information that was, for all practical purposes, consistent with the information that would be derived from the application of more complex and sophisticated models.

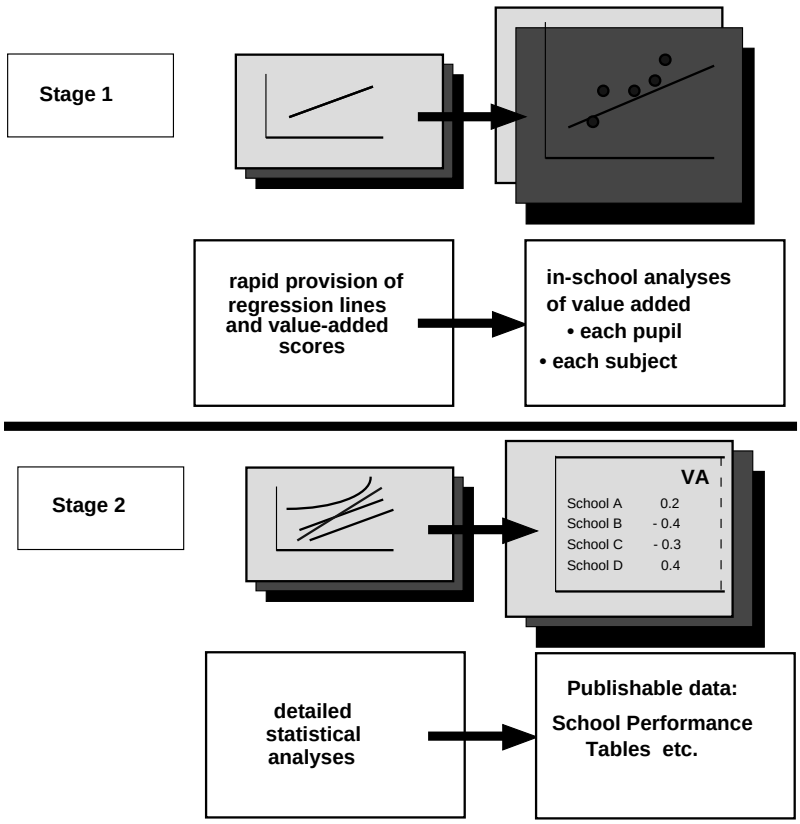
---

[1]Acronyms and technical words are defined in a glossary, Annex G.

1.2.1 A suggested two stage approach

A two stage approach to data analysis was accepted as a possibility. The first stage would respond to the needs of schools for readily understandable and quickly provided data. The second stage would allow for thorough exploration of the extent to which additional analyses were needed in order to be as fair as possible to every school (Figure 1.2). If there were features in the dataset which could be used to make the comparisons fairer, then this work should be undertaken before results were released for publication. Such **exploratory data analysis (EDA)** takes time.

Figure 1.2 A two stage approach to the analysis of value-added data



1.3 Value-added scores defined as relative progress

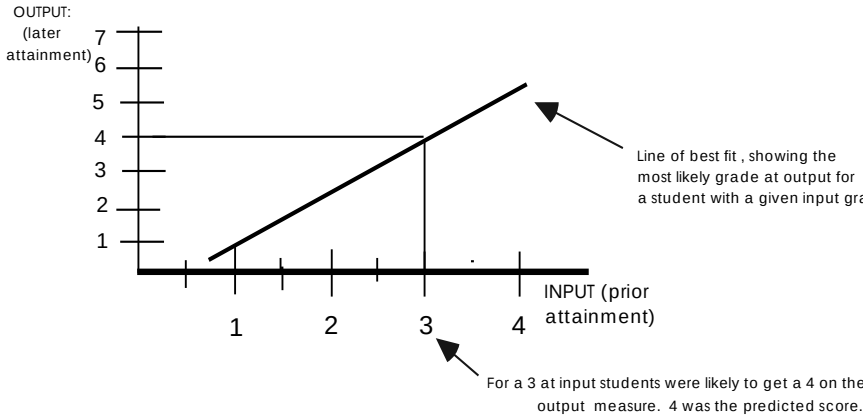
The project defined a value-added score for a pupil as a measure of the progress made by that pupil relative to the progress made by other pupils i.e. relative progress. Simple measures of relative progress were derived by computing, for each year's data, a **trend line** showing roughly how pupils had tended to achieve from their various starting points. For each pupil a 'statistically predicted score' could then be calculated, representing the score that was most likely for pupils starting from the same point. The difference between the score a pupil actually achieved and that predicted, provided the value-added measure for that pupil, a measure of that pupil's progress relative to that of similar others. The statistical term is 'residual' denoting that which is left over after the prior achievement has been taken into account. (as illustrated in Figure 1.3) However, the term 'value-added' is now in such widespread use that this terminology is unlikely to

Figure 1.3 Value-added from 'residual gain analysis' (RGA)

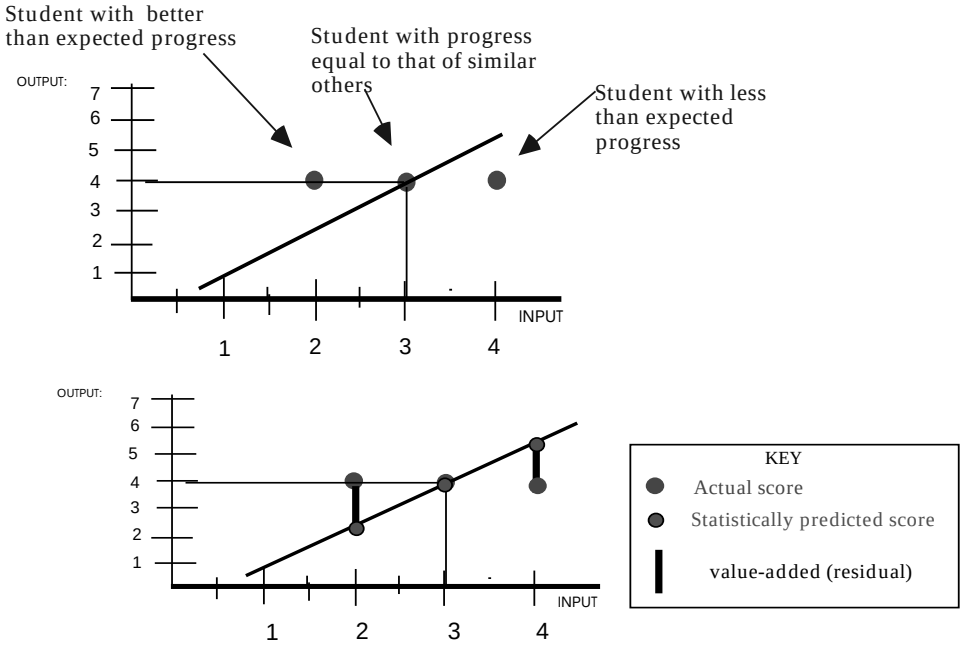
[Note: The term 'score' is used. It could refer to key stage levels , to test marks or to grades.]

1. Nationally representative data are used to produce a graph showing the general trend from entry to exit, i.e. input to output.

2. Using the trend line, each pupil's 'predicted' score can be calculated. This is the score that pupil was most likely to get considering the input for that particular pupil i.e. it roughly represents average performance for similar pupils.



3. The scores pupils actually obtained are compared with their predicted scores. The difference, which is often referred to as the value-added, represents the *progress* made by the pupil *relative* to similar others. If the difference (actual score minus predicted score) is positive then a pupil has made greater progress than the average of similar others. If the actual score is lower than predicted then the pupil achieved less progress than similar pupils.



4. Averaging the value-added scores for a school or class can provide an average value-added measure for that school or class.

change. The 'residual gain analysis' just described can be seen as simply a way to compare the progress of a school's pupils with the progress of similar pupils in other schools.
If a pupil made more than average progress the 'value-added' will be positive, with higher achievement than predicted. If a pupil made less than average progress, the value-added will be negative, with lower achievement than predicted. Average progress is indicated by there being no difference between the statistically predicted score and the actual score. Although this quantity is zero it indicates *average progress* not *no* progress. A school with an average value-added score of zero has kept up with other schools; on average, its pupils have made appropriate progress. The progress made by pupils cannot be entirely attributable to the school, as is implied in the phrase 'the value-added by the school'. The value-added indicator for the school will be approximate, will vary from year to year and will reflect the effects of unmeasured influences that might well be specific to particular pupils. However, this quantity is a simple and clear measure, an indicator that can be regularly monitored. It comes closer to reflecting the relative academic successes of pupils than does any other simple measure. In view of the fact that about half[2] of the variation in examination and key stage testing results can be predicted if the prior attainment of pupils is known, it is widely accepted that fairer comparisons among schools require that some account is taken of their widely differing intakes.
However, much is yet to be learned about the extent to which the value-added indicator can be altered by the efforts of schools.

1.4 The first phase of the Value Added National Project

Results from the readily understandable 'residual gain analysis' procedure were compared with the most sophisticated technique available, multi level modelling (MLM.) a technique that requires specialised software. The term 'multi-level' refers to the levels of data. Measurements on each *pupil* are usually the first 'level' and measurements of *schools* the second level. Measurements related to *LEAs* could be considered a third level. The term 'hierarchical linear model' is used alternatively to 'multi-level modelling' and indicates the **hierarchical** nature of the datasets to which multi-level modelling is appropriately applied: pupils within schools, schools within LEAs. These levels can be considered without the use of specialised software but such software is designed for multi-level data and has advantages for research purposes. (but see Annex C.)

1.5 Major Project Activities and Reports

The statistical trials in the first phase of the project had relied on data already available, and such data was fairly sparse. The second phase of the

---

[2] The typical correlation of 0.7 between a prior achievement measure and a later test or examination result implies that the percentage of the variation in the results that can be predicted from the input measures was (0.7*0.7) = 0.49 or approximately 50 percent.

project was designed to explore further the feasibility of systems, looking again at Key Stage 2 and Key Stage 4 and, this time, running pilot projects. There was concern about the likely acceptance of value-added measures in schools, the nature of the key stage tests, the data that would need to be retained by schools or elsewhere, the methods of reporting to schools, the issues of excused absence, illness, truancy and mobility and whether or not the simple measures would again stand the test of being adequate. Equally the issue of whether a second stage of analysis was needed had to be considered. Above all, if a national system was indeed feasible, how should it be organised?

> *'We need to look at issues of management, to see how feasible it is for schools to collect and keep the necessary information. Any system must be manageable.'*
>
> Sir Ron Dearing in the Preface to the 1995 General Report.

In short, what could be learned about how to run the system as cost effectively as possible? What kind of organisational model might be adopted? What defensive planning was needed in order to avoid future problems? It was to these issues that attention was turned during the second phase of the Project.

The major activities for both phases of the project are listed in Table 1.2. They included investigations of the statistical characteristics of existing datasets, the collection of additional datasets, running pilot projects, surveys, meetings, discussions and experiments on methods of data presentation and support for schools.

Table 1.2 Major project activities and sources of data

Technical report number	DATASET	Date of report	TYPE OF STUDY	No. of Schools	No. of pupils
		1995			
Primary 1	PIPS '95	Dec.	statistical analyses	110	4,776
Secondary 1	'94 GCSE	Dec.	statistical analyses	39	5,209
		1996			
Primary 2	'96 expt.	May	Experiment: types of value-added feedback	104	na
Primary 3	'95 trial(AVON)	July	statistical analyses	247	8,224
Secondary 2	'96 expt.	August	Experiment: types of value-added feedback	389	na
Primary 4	'96 pilot	Nov.	Pilot of a value-added system	396	13,626
Secondary 3	'96 pilot	Dec.	Pilot of a value-added system	185	25,843
Additional sources of information					
	Sept	1995	Round table meeting with statisticians		
	Apr.	1996	Survey of, and Seminar with, LEAs		
	Oct.	1996	Questionnaires from two one-day conferences for pilot schools		
	Dec.	1996	Questionnaire to Secondary Pilot headteachers regarding use with parents and governo		

1.6 Structure of This Report

Figure 1.4 illustrates the way in which sections of this report relate to decisions that have to be made. In phase 1 of the project (represented by decision points 1a and 1b) we came to the conclusion that, somewhat against widely held expectations, the end of key stage test data could be used to predict subsequent achievement with the kind of accuracy generally found in studies of school effectiveness. Furthermore the datasets confirmed the finding from many other datasets: that the use of sophisticated statistical models that were certainly not as 'readily understandable' as a simple procedure, yielded results that were not substantively different from those yielded by simple analyses. Consequently our answer to the question 'can it be simple? ' was 'Yes, for the vast majority of schools, for most practical purposes.'

In either case, whether a simple or complex analysis is used, we move on to section 2 of the report. Will value-added systems be rejected or welcomed? This is an important consideration in that the quality of education delivered depends upon teachers. If teachers would be extremely unhappy with a value-added system this would have consequences for the running of the system and the general morale in the profession. It was therefore very important to establish whether the profession was interested in value-added analyses, would welcome them

and how they would envisage such systems being used with parents, governors and with regard to the issue of their publication, possibly in a revised version of the School Performance Tables. This issue is explored in Section 2 which draws on both surveys and experiments on methods of providing data. The evidence is that value-added systems will be welcomed by headteachers, although there are still misgivings regarding the quality of the tests, the sharing of the data with parents and the publication of data. There is also a clear need and demand for training in value-added concepts, for headteachers, parents, governors and inspectors. The welcome given to value-added was both from those with experience of such systems and from those without such experience. Had the thousands of schools that are participating in value-added systems found the data peculiarly out of line with their experience and therefore in doubt regarding its validity, this welcome would not have been extended.

There seemed, therefore, to be every reason to recommend that a national system should be developed for each key stage and the remaining question was whether the time and expense of a second stage of analysis would be cost effective. Was there sufficient complexity in the data that further light could be cast by employing advanced statistical techniques on large datasets? Section 3 in this report summarises the findings from all the datasets and the pilot studies.

In Section 4 the kind of data that might be provided to schools in a national system is described and uses for such data discussed.

Having shown what a national system might provide the remaining issues were of a practical nature. Section 5 considers organisational models and these inevitably raise philosophical and political questions about how best to run a national system. Our anchor point in such discussions is to design a system as consistently as possible with knowledge of how science advances, and how complex systems evolve. An additional principle is to seek the most cost-effective methods.

The design of a national system that will feed back to each teacher, for every pupil, information on the relative progress that the pupil has made in the preceding key stage, raises a wide range of issues such as, for example:

- the heavy emphasis on cognitive achievement perhaps to the exclusion of other concerns;
- the potential misinterpretation of data with consequent misdirection of time and effort;
- the stresses in a system that can lead to actions designed to alter indicators rather than to alter the reality underneath the indicators;
- the need for considerable safeguards if there is to be publication of data which will be seen as having powerful consequences.

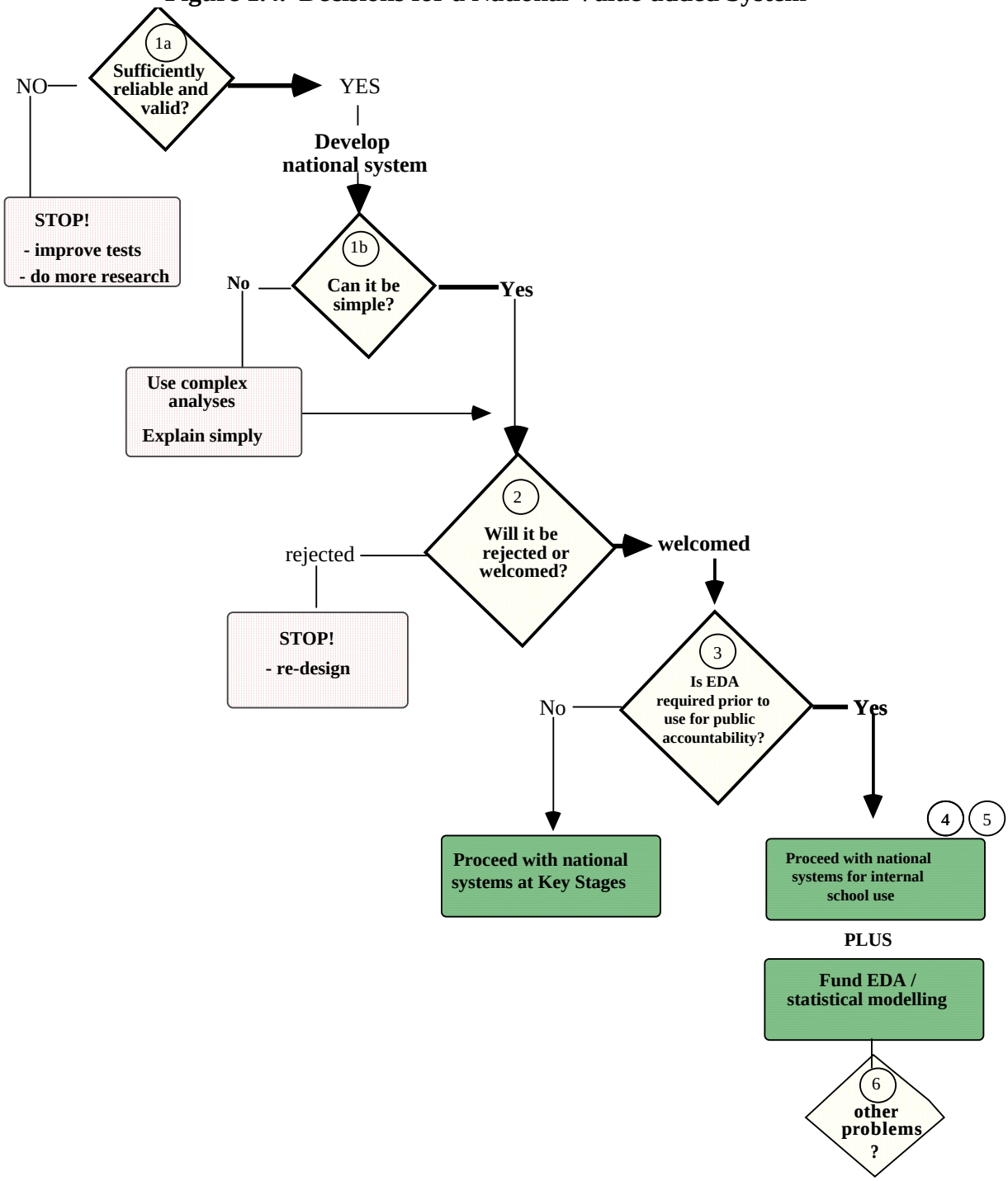
These wider issues are addressed in Section 6, with the aim of attempting to avoid problems by foreseeing them. The report concludes with a series of recommendations, cross referenced to this report and the technical reports.

1.6.1 Annexes

The Annexes to the Report include:

A A list of the reports produced for the Value Added National Project
B A list of members of the Advisory Group
C A discussion of Modelling issues
D Illustrations of error terms in simple and complex analyses(Tymms)
E Estimating size of cohort needed for reliable value-added measures
F The fifty percent framework: extending value-added measures
G Glossary of terms
H References

Figure 1.4. Decisions for a National Value-added System



1 indicates SECTION number in this report

1 = Introduction
2 = Views on value-added
3 = Quantitative Findings
4 = Using value-added information
5 = Organisational models for a national system
6 = Wider issues
7 = Recommendations

SECTION 2 :VIEWS ON VALUE ADDED

If a value-added system is to be effective in supporting high standards in schools, it will need to be both widely accepted and well used. Two sources of influence on acceptance might be the way the data are reported and the amount or type of support provided. We addressed these issues with a number of studies, some going beyond the usual surveys and employing randomised trials designed so that equivalent groups of headteachers responded to different kinds of value-added feedback and assistance. Results are summarised below with regard to the current levels of interest in value-added, the type of feedback, the uses to be made of the data (including publication) and the perceived need for support in schools as value-added systems are introduced. Full reports of these studies are available in the Technical Reports listed in Annex A. Here the major features are summarised and comments from secondary headteachers are included that are not available in the Technical Reports.

2.1 Sources of evidence

Two groups of headteachers, one Primary headteachers in Avon and the other a random sample of secondary headteachers, were provided with value-added data in various formats and asked to complete a questionnaire. Each study had a response rate of about 40 percent from the headteachers who had been approached, yielding 104 primary headteachers (from the 257 in Avon) and 389 secondary headteachers (from 906 randomly selected). The Avon Primary headteachers were answering after they had received their first information on value-added. Thus the data to which they were responding was actually that referring to their own school and, as might be expected, headteachers tended to be more positive if their (real) value added scores were positive.

The secondary experiment used realistic data that was the same for all schools and could not, therefore, account for differences in responses. Secondary headteachers knew that the data to which they were responding was not related to their school.

2.2 Interest in value-added information among headteachers

Among both primary and secondary headteachers there was an overwhelming endorsement on the returned questionnaires of the statement that value-added information, made available confidentially, would be "beneficial" in its impact on the school. (85 percent primary headteachers and 95 percent secondary). Furthermore the vast majority of secondary headteachers (85 percent) rejected the statement that 'Statistics cannot tell me anything that I do not already know about my school". Primary teachers were less emphatic and here a substantial minority (34 percent) agreed with the statement. Among headteachers, in general, quantitative performance data was acceptable, indeed welcome.

This response was further confirmed by additional comments. About 40 percent of secondary respondents took advantage of the opportunity

provided on the questionnaire to write in comments, some brief and some extensive. Although self selected respondents cannot be considered representative of all headteachers, the overwhelming impression was of a profession keen to receive Value-added information but not at all keen to see it published in the press.

Almost universally the use of value added was welcomed for internal management, particularly in looking at the effectiveness of departments from year to year. The need for simplicity was frequently stressed. It seems that the Secretary of State was accurately reflecting the views of teachers when she said that Value Added indicators should be 'readily understandable'.

- *Make it simple and fair and produce a system quickly please!*
- *Would welcome involvement in a project of this nature.*
- *The sooner we have a national system the better.*
- *Encouraging for staff who work with children who are educationally disadvantaged - provides a measure of their work. Equally, shows when teachers are complacent with bright/advantaged children.*
- *Excellent idea. But get it right so schools understand the formatted data long before publication publicly in order to avoid pit falls which can occur.*
- *1. If we must have league tables, value added ones are the only possible fair way. 2. Value added information can be used very beneficially, internally within a school.*
- *I have been involved in various value-added schemes for 3 years now and each year I find the material of greater value and its reliability is steadily increasing.*
- *The concept of value added is an excellent one. The information it generates can be an essential analytical tool when it comes to a school's performance. However, the method of calculating value added has to be a simple easily understood one that staff can operate and governors and parents can understand.*

But dangers were recognised:

- *Must be seen as a way of enhancing what we do - not a stick to beat us with*
- *Sophisticated implications and messages from value-added may be beyond most people and could lead to greater lack of confidence in schools and education system..*
- *Value added comparisons nationally will only be fair if similar schools in terms of socio-economic intake are compared. Hence local comparisons would have greater validity in a first analysis.*

Two of the few comments opposing the development of value-added systems were:

- *Despite the profession's justifiable views on value added, parents are only really interested in the raw results. A school that gets lots of 'A' grades will always seem to them to be better than one which achieves Ds with an intake of lower ability pupils. I do not therefore feel an elaborate system devised to produce national value-added results is value for money.*
- *I cannot really see it running on a national level. It would need a lot of explanation for people to understand it.*

2.3 Responses to the different types of value-added data

Randomly equivalent groups of both primary and secondary teachers were provided with either graphs or tables. For primary headteachers the graphs showed bar charts of actual and expected scores and the tables listed each

pupil's Key Stage 1 score along with the predicted and actual Key Stage 2 score and the difference between these scores i.e. the value-added
Among Primary headteachers there were interesting trends suggesting that feedback in the form of tables rather than graphs was the more successful: a higher response rate to the questionnaire followed the receipt of tables, and, remarkably, the value-added was significantly higher among the next cohort of pupils in the schools of those headteachers who had received tables as opposed to graphs only. It seems possible that tables were better accepted and used than graphs.
Why were numbers more influential than diagrams or was this a chance finding despite its reaching the traditional levels of statistical significance and a moderate effect size of 0.2? Was it possibly the case that headteachers who received tables of numbers shared these with their staff and thus the data began to have the influence that is widely attributed to feedback, including the School Performance Tables: a focusing of effort on measured achievement?
Among secondary headteachers bar graphs representing the performance in 10 subjects were the clear first choice. They were perceived as most useful and the easiest to explain. When the results in the different subjects were combined to create a single, whole school indicator, this information was seen as less useful and less easy to explain, no matter what format was used. For this kind of 'school level' indicator it was not bar graphs but box plots that were rated as the method with which headteachers would be most happy (Please see Figure 2.1 for the ratings of various types of presentation of data and Figures 2.2 and 2.3 for examples of bar graphs and boxplots; all are reproduced from the Secondary Technical Report No.3.). The choice of boxplots for reporting data referring to schools seemed wise since the bar graphs of average scores tend to exaggerate differences whereas box-plots illustrate the large amount of overlap that is present even in datasets with different average scores.
The choice of publication in box plots would be a difficult one to implement in a school performance table. The choice of a simple value-added score would be more likely but would suffer the known problems with ranks: the clustering in the middle ranges so that small differences are associated with large differences in rank-ordered positions. The use of stanine bands or some other form of banding would avoid this but stanines were not relatively popular in the experiment (Figure 2.1) This could be partly because of unfamiliarity and a lack of explanation of stanines in the accompanying text.

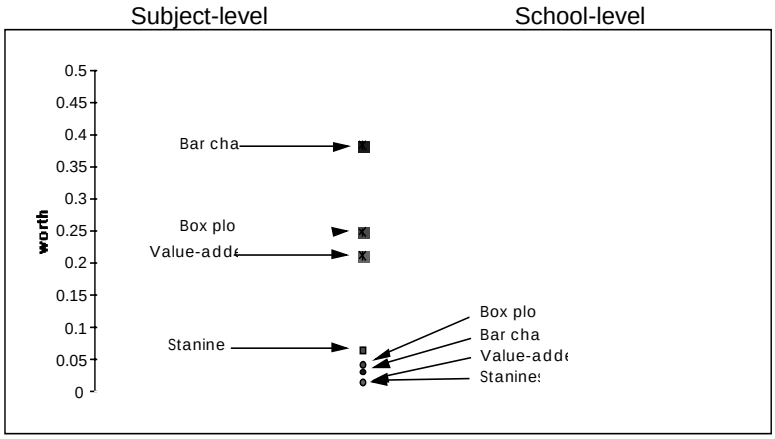
2.4 Attitudes to the publication of value-added data

Whereas secondary headteachers had overwhelmingly seen value-added feedback as beneficial for their school, less than a third saw the *publication* of value-added data as likely to be beneficial. However, given the fact of publication nearly two thirds would want value-added information published alongside GCSE data if that is published.

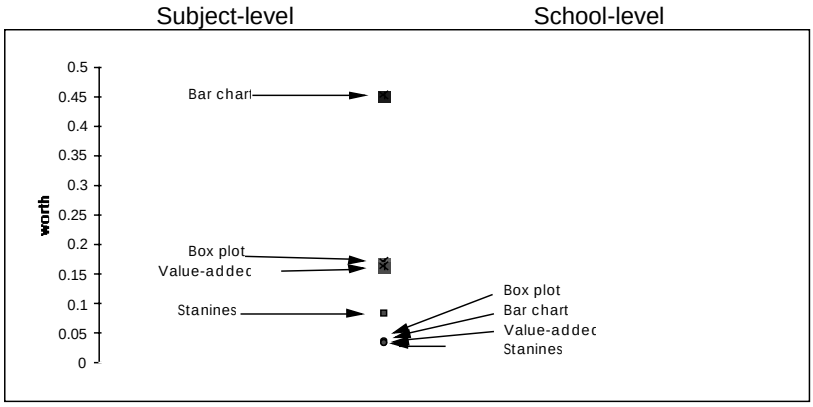
Primary headteachers were even more negative towards publication than secondary headteachers. More than four out of five felt that the value-added information given to the press would be harmful or at least a nuisance and only about a quarter agreed that it should be published alongside Key Stage 2 results.

Figure 2.1 The 'worth'[3] of each type of feedback on three criteria

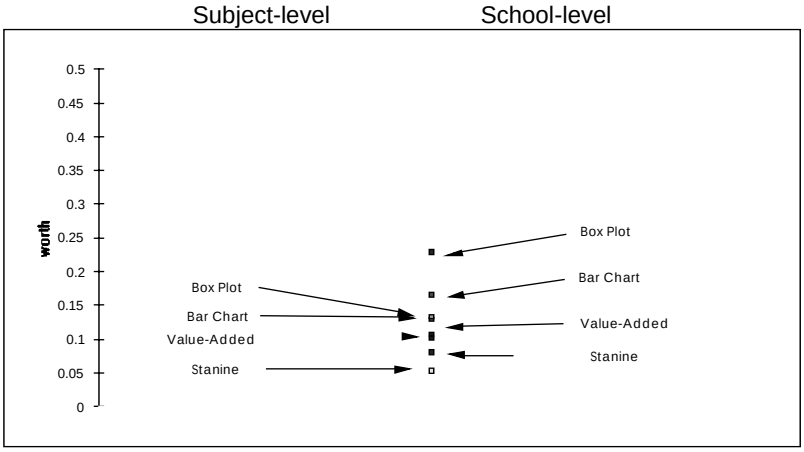
(a) "How useful would you find value-added feedback in this form?"



(b) "Would you feel confident about explaining feedback in this form to staff/governors?"



(c) "Would you be happy to have this information published?"



---

[3]Notes: 'Worth' denoted relative preferences and was assessed by randomly assigned sets of paired comparisons. Please see the Technical Report, Trower and Vincent, 1996) for details of this approach, based on Bradley, R. A. and Terry, M.E. (1952) .

"Value Added" referred to value-added expressed as a numerical score; all data was based on value-added

Figure 2.2 Value-added whole school indicators; bar-chart feedback of ‘school level’ data.

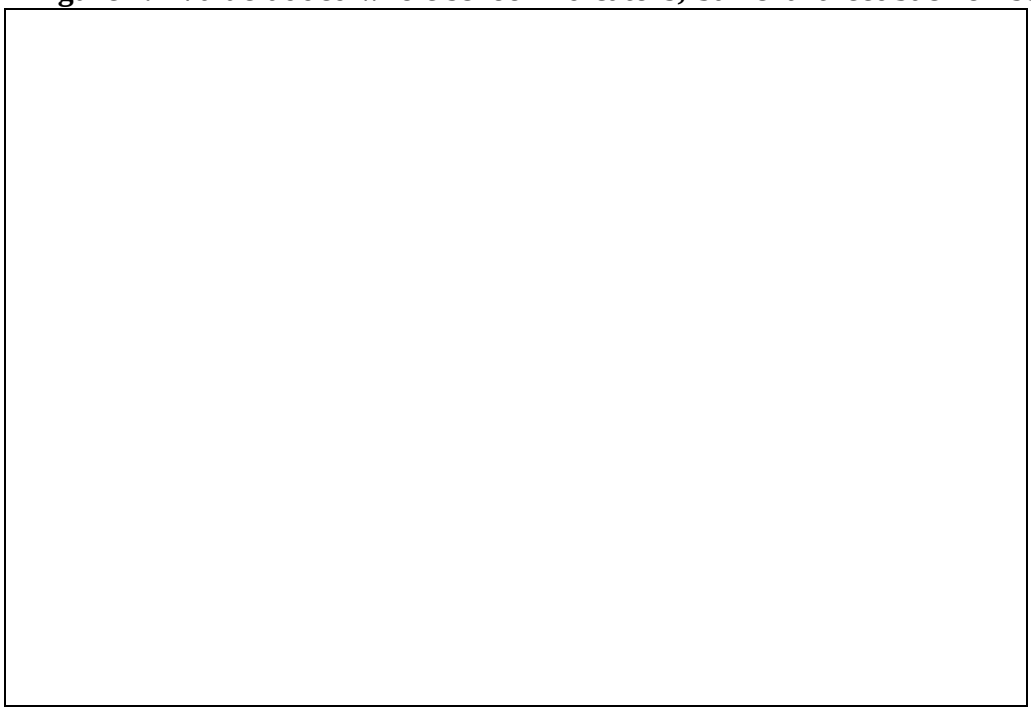
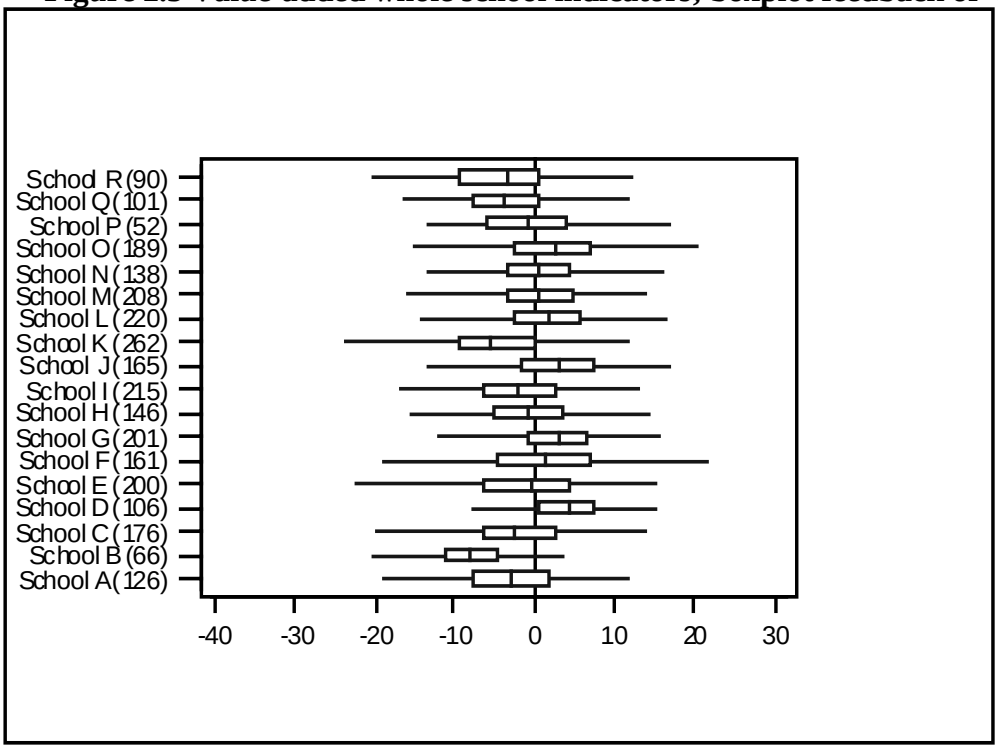


Figure 2.3 Value-added whole school indicators; boxplot feedback of ‘school level’ data.



Representative comments from secondary headteachers are listed in Figure 2.4 and the following is singled out as explicitly indicating problems raised by publication:

- *If the information is to be used inside school we must own it. Publication makes management defensive, so we cannot use it effectively, or complacent so we still cannot use it effectively. It should be a confidential service to schools, but including governors.*

A way forward was suggested by someone less enthusiastic about the value within the school:

- *Publication must be resisted until sufficiently reliable data has been built up over say a minimum of 3 years. It is doubtful how this can help schools improve performance in terms of day to day classroom management.*

Figure 2.4 Comments from headteachers on the issue of publication

Value added does not still make a basis for a fairer system. It is unwise to use it to compare schools which inevitably will be done by the media/parents etc. who are not knowledgeable about education or understand that every school is unique and there is more to education than academic results.

I prefer it to a straight publication of GCSE scores, but any figures that get into the press will be viewed in success/failure league tables. Any effective value-added system will be better than the current situation but the current situation is dreadful!!! The publication of league tables itself is the problem

We make quite extensive use of value-added information, particularly in assessing performance at GCSE. It is used internally and diagnostically. I have serious reservations about publishing value-added data because such a lot of other information and understanding about the school is essential in order to interpret it correctly and to draw valid conclusions from it.

1. We feel well advanced with our school's value-added project. 2. I don't trust our local press to deal fairly with any statistics/information. We would like 'ownership' of the information so we could manage the form/timing of its communication to the press.

National system = loss of control of information from the school = danger of usage for crude comparisons as opposed to positive intervention with the school.

There ARE REAL problems for the 40+ per cent of schools/departments which, inevitably fall below average. I remain unsure that the publication of all this is definitely helpful to them.

I think the use of value-added statistics is a valuable management tool and can help to inform individual parents. However there are too many assumptions and variables for it to be used in published results e.g. a secondary schools results may vary according to the strength of their primary schools. I believe therefore the raw scores are the most helpful way of publishing exam results.

Fiddling the data in high-stakes systems

A system in which publication makes the data highly consequential for schools is often referred to as a 'high stakes' system. Concerns were expressed that, in such a system, the use of prior attainment was, perhaps,

not the best way forward for value-added measures because of the temptation to depress an intake measure.

> *A model based on KS3 testing is flawed because, if I want to enhance the school's value-added performance all I have to do is to maintain concentration on GCSE performance whilst taking a less than rigorous approach to KS3 testing which does not have the same profile. A more effective comparison would be achieved using KS2 test results.*

This concern was also raised in the meeting that the project held with statisticians and indicated by the following comments transcribed from the audio-taped proceedings:

> *If you are worried about measurement errors you then have to consider the process that generates the data. ...At, say, primary level with school based assessments at KS1 and KS2, then if one school produces both assessments the process is different from that where one school produces the KS1 score and a different school produces the KS2 score. The incentive to fiddle the results is quite different in the two cases.*
>
> Professor T.M.F. Smith, University of Southampton

The possibility that such an effect had already been seen in the data was mentioned.

> *I think there are certain problems to do with using high stakes...- the outputs from one stage and the input to the next. ..It's better to use as an input something which....is just there to test where pupils are at this stage, which nobody has any particular interest in - - - if I say 'fiddling' it might sound pejorative - but in no particular interest in having anything other than the truth at that stage and preferably something which is finely differentiated and which preferably is the same from year to year. ..... I mean, that's my ideal: a value-added system is something which has a consistent, not high stakes, finely differentiated input measure year after year.'*
>
> Ian Schagen. National Foundation for Educational Research

Headteachers also saw a need for more fine-grained measurements

> *As a grammar school our (end-of-KS3) results are compact around level 6-8 and I wonder how useful they are as a base for value added!*
>
> *Many of the pupils arriving at our school in Year 7 have low scores in reading and numerically. This particular comparison KS3 results with GCSE score represent value added over the two final years of schooling. To be really beneficial the value added should be taken over 5 years year 7 - KS3 - KS4 end.*

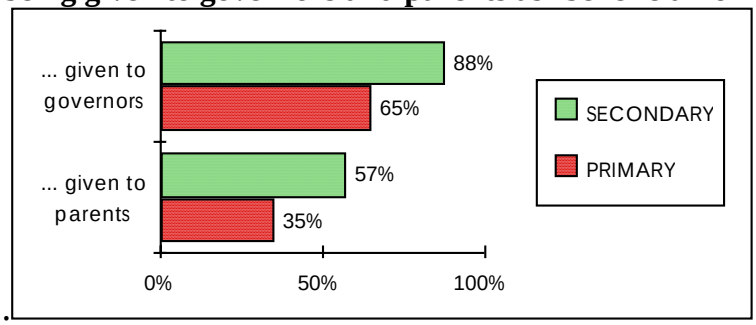
At secondary level the wisdom of looking only at the last two years of compulsory schooling (Key Stage 4) was questioned:

> *If we do really well with pupils at KS3 and they achieve good results, the value added at KS4 won't really be very much (if any) and the school will appear to have done badly between KS3 and KS4 whereas really it did very well at KS3.*
>
> *If most of the improvements have occurred in KS3 it will not appear in subsequent measurements.*
>
> *Currently Year 9 pupils do not take (end of key stage tests) too seriously - results are often not representative of true ability.*

2.5 Use of value-added data with parents and governors

Both primary and secondary headteachers were generally positive when asked to consider value-added information being given to governors, seeing the impact as probably beneficial. There was less certainty with regard to the use with parents but for neither group were headteachers opposed. Questionnaire responses are summarised in Figure 2.5

Figure 2.5 Percentage of headteachers who rated the probable impact of value-added information being given to governors and parents as 'beneficial' or 'very beneficial'



In the secondary pilot, some weeks after full feedback of the school's value-added performance in GCSE for 1996, headteachers were asked, by an open-ended questionnaire, if they had used the information with parents or governors, and if so with what effects.

The data had been used with governors by 43 percent of those responding although in almost half these cases it was reported as used in detail only with the curriculum sub-committee. However, most other headteachers reported intentions to use the data with governors which suggested a positive view of the use of such data.

> *Governors found it useful and welcomed the whole notion of value-added as against raw results. They seemed impressed with some of the uses to which it might be put within school.*
>
> *Most well received by governors. They are beginning to understand the nonsense of the 5 A* - C standard and realise the need for good input data at Year 7.*
>
> *It does help to put our modest raw results into context and they can see we add a lot of value.*

with governing bodies, although there were occasional warning signals in the written comments:

> *They thought it useful but it did trigger a rather judgmental approach: what are we doing to "the less successful areas".*

Only 20 percent had used the information with parents and the few comments indicated wariness, as for example:

> *We need to gain more confidence and awareness as a staff before parents could possibly be involved. We do intend to get them involved but that is our phase 2.*

Searching the comments for references to parents led to the location of further concerns with regard to keeping the system simple and also the need for training:

> *Keep it simple - parents will not understand complicated information books.*
>
> *The concept of value added is an excellent one. The information it generates can be an essential analytical tool when it comes to a school's performance. However, the method of calculating value added has to be a simple easily understood one that staff can operate and governors and parents can understand.*

2.6 Advice from headteachers

Since nothing is so credible to headteachers as the views of fellow headteachers, a final question was: "How would you advise other headteachers to use value-added results?" Responses advised caution and wide consultation, beginning within the school and with small amounts of data. The following advice seemed grounded in experience:

1) Invest considerable time in preparing staff (with some governors in attendance)

2) Do not release figures to governors or parents before proper professional analysis
3) Do not use as a reason on their own to change policy/methods - always link to other school data and experience
4) Use a second "back up system" for any project that is in its infancy either in the number of schools/pupils it uses and/or in the time it has run nationally and for one's own school

Again concern was expressed about 'fiddling' the Key Stage 3 data.

Not in a draconian way. Allow curriculum manager to direct departmental support to improve performance. Also allows whole school initiatives to be implemented based on irrefutable evidence. Will it cause some headteachers to try and lower key stage 3 performance if key stage 4 Value Added is published ?

Figure 2.6 Headteachers' advice to other headteachers on the use of value-added data

Work it through with staff first and arrive at your school's policy via consultation and explanation.
Very slowly! It takes about 2 - 3 years for staff to get to grips with it. Don't give out too much data to start with and once you get into it don't over interpret it.
You need a senior member of staff to be very familiar with it.
1) Build up a picture over three or four years
2) Share widely with all heads of departments and members of departments
3) Stimulate discussion regarding predicted grades/actual grades
Avoid making own judgements (at least publicly) - involve those responsible in considering reason for + and - results, ensure they are seen in whole school context - get in department agendas and discussion about quality of teaching and learning - give ownership of data to those who produce the outcomes.
Carefully. Very good internal professional document that allows lots of questions to be asked. Patterns need to be seen over 2 or 3 years before making judgements. It has considerable value.
To bring these to prominence beyond the 5 A - C standard. However I realise that headteachers in "better" academic catchments may not wish to undermine their positions in the irrelevant league tables.*
Talk in whole school terms - avoid department labels as it can be self fulfilling with parents pressure.
Cautiously.
Cautiously - as there is only one year's data!
We have found the information interesting but intend to proceed with caution. However we have found the value added GCSE to A level useful in monitoring pupils who have not reached potential and developing strategies for the next cohort.
(a) to get a better picture of how well the school and its departments are doing
(b) to use as mechanisms for target setting for both departments and individual pupils

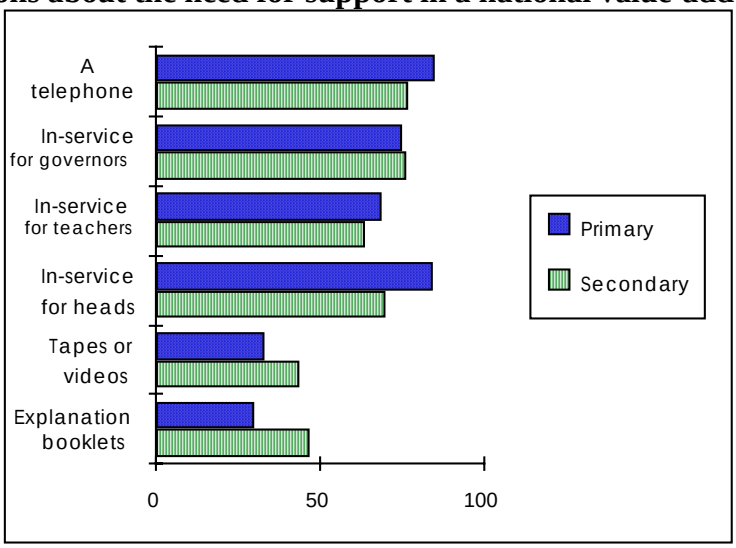
2.7 Perceived needs for support

In both the experimental studies headteachers were asked to consider the support that might be needed if a value-added system were to be run nationally. The questions and the percentage choosing each response are shown in Figure 2.7

Although some headteachers felt that booklets, tapes and videos might be sufficient, the majority saw a need for in-service for headteachers, teachers

and governors and there was very strong support for a telephone help-line for the interpretation of value-added information. Secondary headteachers were asked if 'Local (LEA) comparisons would be useful in addition to the national ones' and 90 percent thought these would be useful.
These responses suggested a preference for person-to-person help, either through in-service or via a telephone help-line.

Figure 2.7 Percentages of primary and secondary headteachers responding to questions about the need for support in a national value-added system.



Note: QUESTIONS IN FULL:

- Explanation booklets would be all the explanation required for schools
- Tapes or videos could provide sufficient training for schools
- In-service would be needed for headteachers
- In-service would be needed for all teachers
- In-service would be needed for governors
- A telephone number to call for help in the interpretation of the Value-added information would be essential

2.8 Experimental evidence regarding the value of INSET

In the primary experiment in Avon, three kinds of Inset, three forms of support (Inset, cassette and booklet) and three ways of representing error were trialled with randomised groups. The only substantial differences arose from the invitation to the Inset. Of those invited (of whom the vast majority attended) a larger proportion returned the follow-up questionnaire, perceived the value-added information as beneficial for their schools and supported the whole value-added exercise. It is interesting to speculate as to whether the experience of Inset was sobering in that a larger proportion also saw Inset as more necessary than did headteachers who had not received Inset.

2.9 Summary and conclusions

The picture that emerged was one of a profession that welcomed readily understandable statistical data relating to the relative progress of pupils. Considerable numbers of headteachers were already benefiting from participation in informal value-added systems These informal systems lack the stamp of official approval, lack the coverage that a national system would make possible and require funding by the school or the LEA, but they have carried the advantage of confidentiality, enabling schools to

become familiar with the information provided by value-added analyses. The time would now seem right to extend the benefits, economically and effectively, to all schools.

The experiments reported in this section give some indications of the kind of data appreciated by schools and the perceived need for Inset. The main concerns centred on the possible temptations to 'fiddle' the data by depressing intake scores, the limited value of a single key stage and, above all, the publication of value-added data.

It is difficult to oppose publication in an age of freedom of information although the concerns of headteachers are undoubtedly well founded and have been observed in other systems of performance monitoring, as indicated in Section 6 of this report.

Among possible ways forward are

- the publication of value-added indicators of some kind in the School Performance Tables. At the secondary level this was favoured by a majority of the secondary headteachers if there were to be School Performance Tables (Section 2.4 above).
- a forewarning that publication would be probable in a matter of about three or four years' time when three years of data could be averaged to provide more stable indicators and sufficient numbers of schools would have retained the appropriate intake data
- voluntary publication, with schools participating to the extent that they wish to do so.
- publication of the names of schools reaching the top 10 percent of schools in each of the major subjects. This solution would possibly be highly motivating to school departments, any of which might aspire to this recognition. It would also facilitate bench marking.

In Section 3 the findings from the quantitative analyses will be summarised, with particular attention to the implications for the methods of running the national value-added system that seems, now, so appropriate.

SECTION 3: QUANTITATIVE FINDINGS

The aim in this section is to bring together the weight of evidence that underpins some of the recommendations made in Section 7. The evidence derives largely from the data analyses and pilots reported in detail in the five technical reports. The 'findings' are numbered for ease of reference. They are, of course, findings limited by the datasets available and their generalisability to other datasets will need on-going investigation.
The evidence is grouped according to the following questions:

Question A. Is there a basis for a statistically valid and readily understood national value-added system in the near future, for use in internal school management and professional evaluation? (Findings 1 through 7)

Question B. Would a second stage of data analysis be necessary prior to making value-added indicators public on a school by school basis? (Findings 8 through 10)

Question C. Are there benefits to be gained from the inclusion of additional variables such as sex, social indicators, mobility measures, attendance data and examination syllabuses/boards...
(a) for internal school management?
(b) for publicly available information? (Findings 11 to 15)

3.1 Question A. Is there a basis for a statistically valid and readily understood national value-added system in the near future, for use in internal school management?

The five datasets that were described in Section 1 are listed in Table 3.1 below along with the data needed to show that value-added systems are viable with currently available tests and examinations (findings 1 and 2 below.)

Finding 1. The end of key stage tests were adequate predictors of subsequent performance.

If the currently available measures of prior achievement did not correlate highly with subsequent achievement there would be little point in looking at the relative progress pupils made, ie their 'value-added'. The correlations obtained over the last few years can be seen in column 7 of Table 3.1 below, and were acceptably strong, with the possible exceptions of primary science and some subjects at GCSE. In column 8 the proportion of the variation in the outcomes that could be predicted from knowing the

prior attainment levels is reported[4]. The prior measures generally 'explained' about half of the variation in subsequent test or examination performance, slightly less for some of the individual subjects in the primary sector and more for most of the secondary measures. Since prior attainment levels generally accounted for much of the variation in raw outcomes, there was clearly a basis for value-added calculations.
The quality of the input and output measures will affect the reliability, as well as the credibility, of any value-added system and it was clear from written comments from head teachers that the end of key stage tests had been a matter of concern. However, these tests are improving and it was clear that they already served as adequate predictors of subsequent performance in that they yielded data with relationships as strong as usually found in research on school effectiveness. The end of key stage tests, then, could be considered sufficiently reliable for the purpose of being used in value-added systems. Work on direct measures of their reliability is on-going within SCAA.

Finding 2. The best basis for prediction was provided by the average of prior attainment measures.

To predict English from English and mathematics from mathematics might seem reasonable but, empirically, it is not the most effective use of prior attainment data. It can be seen in Table 3.2 and 3.3 that the 'predictor' used in every dataset was an *average* of such prior attainment data as was available. This was because, in every dataset, the average measure produced either the highest correlation with subsequent achievement in individual subjects and with overall achievement or a correlation scarcely different from that provided by the best single predictor. This finding is predictable from classical test theory regarding the reliability of tests, which shows that longer tests will be more reliable, other things being equal, than shorter tests. By averaging several measures the equivalent of a longer test is obtained and more information is used.
Since the average of prior attainment measures produced the best basis for prediction of each subject, a single figure kept on record for each pupil would provide the basis for value-added calculations in all subjects. This indicates a minimal requirement for record keeping to enable value-added measures to be computed.

Finding 3. Using marks rather than levels had little effect on value added indicators.

For each dataset school value-added indicators were computed using marks and using levels. The results correlated highly (0.98 in the two secondary datasets and all above 0.92 in the primary datasets.) However, for schools with extremely low or high intakes, marks might make a difference and for internal analysis pupil by pupil, a finely differentiated scale would be

---

[4] This is simply the correlation coefficient squared.

appreciated. The benefit of levels is that they can be directly interpreted each year, as can GCSE grades.

Whilst an average across subject levels provided a sufficiently good *input* measure for the purpose of value-added predictions in the datasets available, more finely differentiated *outcomes* would be welcome. The use of supplementary tests, such as curriculum-free tests of aptitudes, can be used to obtain finely differentiated *input* data, and to cast additional light on the performance of individual pupils, a topic addressed briefly in Section 4, but it will be important that outcome measures differentiate across the entire scale, with no ceiling or floor effects.

Table 3.1 Key data relating to the reliability of value-added systems with currently available tests and examinations.

1	2	3	4	5	6	7	8	9
DATASET	No. of Schools	No. of pupils	Average size of cohort: Mean (SD]	Predictor	Outcome	r = corr-elation pupil level	r-square	RHO [1] with control for intake
PRIMARY 1 PIPS '95	110	4,776	45 (33)	Concurrent developed ability[2]	1995 PIPS [3]			
					average	0.66	0.44	.18
					Reading	0.66	0.44	.13
					Maths.	0.70	0.49	.18
					Science	0.62	0.38	.10
PRIMARY 3 '95 Trial (AVON)	247	8,224	33 (20)	1991 Av. KS1 Ma+En.	1995 KS2 av.level	0.68	0.46	.30
					Reading	0.63	0.40	.15
					Maths	0.62	0.39	.22
					Science	0.52	0.27	.28
PRIMARY 4 '96 pilot	396	13,626	34 (22)	1992 Av.KS1 Ma+En	1996 KS2 av.level	0.74	0.55	.22
					Reading	0.67	0.45	.15
					Maths	0.68	0.46	.14
					Science	0.62	0.38	.18
SECONDARY 1 '95 analyses	39	5,209	134 (49)	1992 Av.KS3 [4] (Ma.+Sci)	1994 Av.GCSE	0.78	0.61	.17
	24	2731		"	English GCSE	0.66	0.44	.09
	24	2682		"	Maths. GCSE	0.78	0.61	.18
SECONDARY 3 '96 pilot	185	25,843	139 (58)	1994 Av.KS3 (Ma+ En.+ Sci)	1996 Av.GCSE	0.87	0.76	.12
					Total GCSE	0.83	0.69	.15
					English. GCSE	0.77	0.59	.12
					Maths.GCSE	0.85	0.72	.11
			various		28 other subjects	from 0.50 to 0.80	from 0.25 to 0.64	

NOTES:

Av. = average or mean En = English Ma = mathematics Sci = Science.

[1] Rho = 'Intra-school correlation, a measure of the extent to which there was variation between schools, i.e. the extent to which schools were different on the measure indicated. (Further explanation in the glossary.)

[2] Specially developed Science and mathematics tests based on the National Curriculum, and a published, standardised reading test.

[3] Developed ability as measured by a PIPS project test.

[4] English KS3 was not introduced until 1993

Table 3.2 Correlations between end of Key Stage 3 test scores and GCSE grades : secondary pilot

GCSE	KS3 English	KS3 Mathematics	KS3 Science	Average KS3	No. of pupils
Average GCSE points score per subject entry	0.75	0.80	0.77	0.87	14,281
Total GCSE points score	0.72	0.76	0.74	0.83	14,281
Art & Design	0.49	0.48	0.48	0.55	4857
Biology	0.50	0.50	0.61	0.64	1144
Business Studies	0.58	0.63	0.60	0.71	2736
Chemistry	0.49	0.53	0.61	0.65	1118
Classical Subjects	0.47	0.49	0.44	0.57	279
Combined Design & Technology	0.62	0.63	0.58	0.68	406
Comb. Information Technology	0.50	0.54	0.53	0.61	1103
Computer Studies	0.54	0.56	0.53	0.63	307
Craft, Design & Technology	0.54	0.56	0.57	0.63	3114
Design & Technology	0.61	0.66	0.62	0.70	2148
Economics	0.52	0.58	0.57	0.66	299
English	0.74	0.67	0.64	0.77	14,004
English Literature	0.68	0.61	0.59	0.71	12,002
French	0.66	0.65	0.61	0.73	8297
Geography	0.67	0.71	0.70	0.79	7355
German	0.62	0.62	0.58	0.70	3753
History	0.65	0.67	0.66	0.75	6007
Home Economics	0.58	0.56	0.57	0.65	6482
Information Technology	0.61	0.59	0.54	0.67	410
Mathematics	0.63	0.85	0.76	0.85	13,816
Music	0.57	0.56	0.53	0.62	1071
PE / Dance	0.50	0.58	0.59	0.63	1711
Physics	0.43	0.56	0.61	0.64	1189
Religious Studies	0.65	0.60	0.60	0.70	3368
Science: Double Award	0.60	0.73	0.76	0.80	10,637
Science: Single Award	0.50	0.65	0.67	0.71	888
Social Science	0.56	0.58	0.60	0.67	254
Spanish	0.63	0.63	0.58	0.71	1251
Technology	0.45	0.40	0.41	0.50	451
Welsh	0.69	0.53	0.47	0.62	103

NOTES: The data were from pupils with results from all three end of key stage tests. (The strongest correlations have been highlighted.)

Table 3.3 Correlations between end of Key Stage 1 test scores and Key Stage 2 levels : primary pilot

Key Stage 2	KS1 Reading	KS1 maths.	KS1 Science	Av. KS1 Engl +maths [1].	Av. KS1 Engl+ maths+ science
Average Key Stage 2 level	0.65	0.63	0.52	**0.68**	0.67
Key Stage 2 reading	0.63	0.56	0.46	**0.64**	0.61
Key Stage 2 mathematics	0.59	0.60	0.47	**0.63**	0.61
Key Stage 2 science	0.51	0.51	0.44	**0.54**	0.54

Note: [1]There appeared to be a problem with KS1 science such that the English and mathematics alone gave slightly better prediction.

Finding 4. In primary schools, value-added scores for individual subjects correlated quite strongly whereas in secondary schools intercorrelations among subjects were low, suggesting a need for a value-added profile for each secondary school.

When value-added is measured pupil by pupil, the data can be summed to provide a single indicator for the school. In primary schools, a single measure across the core, compulsory curriculum would be a reasonable indicator if a single, whole school indicator were required since the indicators correlated highly. However, this situation might change if, for example, subject specialists were introduced. However, in secondary schools whole school figures hide considerable variation from subject to subject. By way of illustration, the correlations between a total outcome measure and English and mathematics value-added indicators are listed in Table 3.4

Table 3.4 Correlations between value-added (VA) measures in English and mathematics and a single whole-school value-added measure

	Primary 1 (PIPS)	Primary 2 (Avon95)	Primary 4 (Various LEAs)	Secondary 1	Secondary 3 Pilot
English (reading) VA with total VA		0.87	0.87	0.73	0.42
mathematics VA with total VA		0.89	0.89	0.66	0.38
English VA with maths.VA	0.77	0.68	0.68	0.49	0.37

Note: These correlations were from multi-level models

With only three subjects in primary schools it was not surprising that two of them correlated highly with total value-added measures based on all three. At secondary level the correlations were lower and schools would therefore look quite substantially different according to which outcome measure was chosen.

In secondary schools there are further difficulties if a choice of a single indicator is desired. A single measure of the five core curriculum subjects would be assessing only about half the work of the school as most pupils appear to take 8 to 10 subjects (mean=7.3; mode = 9). To ignore all the other subjects would risk distorting the attention paid to them.

Furthermore, even if the subjects that were to count was agreed, there would be two justifiable choices for a single outcome measures: the *average* value-added or the *total* value-added. These would not be equivalent unless pupils all took the same combination of subjects. The use of an average would indicate the general quality of the progress made but could be unfair in cases in which only a few subjects were examined. The use of a total would reflect work in all subjects and would reflect, therefore, the cost and effort made by the school in teaching the subjects. Use of a total score is not, however, ideal as it would confuse quality with quantity. It would possibly lead to pupils being put in for as many examinations as possible to boost the indicator.

In short, secondary schools will look very different according to which subjects or combinations of subjects are under consideration.

At the meeting with statisticians in September 1995, the view was very firmly expressed that single, whole school indicators were inappropriate, as indicated by the following transcription :

> *What we're doing at the national level is simply auditing performance, value-added performance.... we're auditing a complex system. And any auditor does not come up with one number which says "Yes, this is fine"... or ..."No, it's a disaster going on here".... I do have professional and moral objections to producing a single summary for accountability purposes. It is not useful at any level. It actually defeats the purpose of trying to produce a better education system, surely..... I think it's the subject level data that tells the story.*
>
> Tony Robinson, NCER & the University of Bath.

Finding 5. A readily understandable model (Residual Gain Analysis) yielded statistically valid indicators of value-added

As was reported in the first two technical reports published in December 1995, the value-added indicators produced by the simple procedure of comparing pupils' performance directly with the performance of similar pupils, regardless of school attended, and then summing the value-added scores (residual scores) gave indicators that correlated so highly with indicators from more complex models that the simple methods could be recommended, particularly to meet the criterion of 'readily understandable'. This issue is taken up at greater length in Annex C but is not particularly important since either approach could be used. The method of determining a pupil's relative performance from a single regression based on the data from all pupils is referred to in this report as 'Residual Gain Analysis' or RGA.

Finding 6 There was considerable variation from year to year in value-added measures

Even using an advanced statistical technique that will increase the apparent stability of findings since the results are weighted towards average, the correlations between one year's data and the next were low. For example, using data from the Avon primary schools, it was found that correlations from 1995 to 1996 were about 0.5, indicating substantial variations from one year to the next, with only 25 percent of the 1996 variance predictable from the 1995 value-added indicators. (Primary Technical Report 4 , p 26) The secondary project did not have datasets from the same schools but data from the Year 11 Information System (YELLIS) was used to illustrate the (in)stability of value-added data across three years. Of the correlations found between the value-added indicators over the years 1994 to 1996, for 43 schools (more than 6,000 pupils), none was higher than 0.74, explaining about 55 percent of the following year's variance. Table 3.5 shows that correlations (the stability measures) were higher for mathematics than for English and Double science.

This variation from year to year may simply be inherent in such a complex system as schooling. Certainly headteachers learn to be cautious.

It must be stressed that .....drawing conclusions from one year's results could be very misleading.

Headteacher responding to the secondary survey.

Headteachers also expressed concern lest others were less cautious:

..... as v-a can swing from year to year (and sometimes for v.good reasons) will parents/governors/press. Public/politicians read too much of significance into it?

Headteacher responding to the secondary survey.

Table 3.5 Stability measures: two years' primary data and three years' secondary

Value Added National Project data			Data from the Year 11 Information System (YELLIS) (n= 43 schools)		
Primary English	1996		Secondary English	1995	1996
			1994	0.44	0.56
1995	0.44		1995		0.53
Primary mathematics	1996		Secondary mathematics	1995	1996
			1994	0.62	0.62
1995	0.59		1995		0.74
Primary science	1996		Secondary (Double) Science	1995	1996
			1994	0.50	0.50
1995	0.49		1995		0.69

Note: All value-added scores were based on multi-level models and the primary ones included varying slopes and were restricted to samples with at least 30 pupils.

Finding 7. With currently available test and examination data reasonably reliable[5] indicators of value-added could be computed for most secondary schools and for about half the primary schools (those with more than about 30 pupils in a year group).

Adequately reliable *tests* are one essential component of a reliable system for measuring value added but we also need some assurance that the *value-added indicators* are sufficiently reliable to be reported. Unreliable indicators would have large margins of uncertainty associated with them so that they would carry the danger of being misleading. Because a value added score might be used for important judgements, the reliability should be as high as possible. A provisional rule of thumb might be a reliability of at least 0.9, the kind of reliability sought in psychometric testing of individuals.

The reliability of the indicators will depend partly upon the number of pupils in the cohort. Small groups will have less reliable indicators than larger groups, other things being equal. If the data is going to be used mainly for internal school management, alongside other information, the reliability becomes less of an issue. Given pupil by pupil data, it is readily apparent if the data are heavily influenced by a few pupils and the variation from year to year will rapidly become apparent, as was indicated by some

[5] 'reliable' as used here refers to the sense that the indicator is reflecting a clear pattern in the data. It is similar to the internal consistency of tests. Table 3.5 reported the stability of Value-added indicators from year to year.

of the comments in Section 2. For publication, however, there would need to be decisions about the size of year group (cohort of pupils) that was acceptable for purposes of publication. Thirty is often advised as a rule of thumb but a more sophisticated approach can be used to vary the cut-off according to other features in the data. The extent to which differences between schools in their average value-added scores can be reliably measured relates to two measures of variability: the differences between the school averages and the variation in pupils' value-added measures within each school. The larger the former is in comparison with the latter, the better the differentiation between schools.

In Annex E an approach resting on statistical models is presented and applied to the five datasets analysed in the Value Added National Project. This method could be applied each year so that the group sizes needed could be adjusted according to that year's patterns in the data. An advantage here would be that the cut-off score would vary. A more arbitrary and stable cut-off score would run the danger of providing a target to avoid. However, a more important point is that the analyses will help to identify subjects that appear to be particularly sensitive to the effects of teaching quality. The methodology will also be helpful in detecting poor examining practices in which, by some means, the school has a disproportionate effect on the outcomes (as is usually the case when internally assessed outcomes are used).

3.2 Question B. Would a second stage of data analysis be necessary prior to making value-added indicators public on a school by school basis? (Findings 8 through 10)

The simplest method for computing value-added scores, one which is readily understood and easily accomplished with almost any statistical package, is to construct a regression line without reference to schools at all, simply using all the available pupil data. (As explained in Figure 1.3, Section 1)

Using only the same data, further analyses can be conducted. These additional refinements can all be made without the inclusion of any further information, ie. without using additional variables such as sex, or socio-economic status. The analyses can also be conducted using either standard regression procedures or one of the advanced packages for multi-level modelling.[6] The advanced packages have been designed specifically to

---

[6]Longford, N. (1990) VARCL. Software for Variance Component Analysis of Data with Nested Random Effects (maximum likelihood). Educational Testing Service, Princeton, New Jersey

Bryk, A.S. Raudenbush, S., Seltzer, M., and Congdon, R. (1988) An introduction to HLM: computer Program and User's Guide. University of Chicago

represent the kind of data available when pupils are measured within larger units such as classes or schools. When there is such 'multi-level' data (eg pupil-level data, classroom level data and school-level data) the packages are appropriate and efficient. The data analyst has to weigh this against the inconvenience of acquiring and using a separate, rather than standard, package. For professionals, this is no problem but school-based data analysts should not feel deterred from looking at data with the procedures available in ordinary packages. Many procedures, such as fitting curves or separate regressions for sub-groups, can be accomplished in standard packages but may give answers slightly at variance with the specialised packages that take account of the extent to which pupils within a group are more alike than a random sample of pupils. The technical reports show that, for the datasets analysed in this project, the simplest data analysis procedure, Residual Gain Analysis generally gives answers that differed only insubstantially from those provided by multi-level modelling . Whatever methods are used, the practical question is the extent to which further analyses, such as provided by a second stage, are needed. It is to that issue that we turn now.

Compositional effects, curves and variable slopes.

In a second stage, the impact of the *composition* of schools, departments or classrooms can be investigated, curves can be fitted, and the regression lines within schools can be separately computed with slopes particular to each school ('variable slopes'). None of these analyses needs further data. Does the move away from the very simplest model yield important benefits in terms of insight into the data? The effects obtained with the available data can be seen from Table 3.6. Columns 2 to 5 indicate the features included in the model. If there is a dot in the column, that feature has been modelled. The rows with no dots represent analyses based on raw data with no adjustments. The simple regression model, taking account only of prior attainment is the next row for each data set. All models used the 'MLn' package.

Columns 6 to 11 in Table 3.6 show the results from the models in terms of changes in the information made available. Thus column 11 shows the correlation of the value added indicators obtained from more complex models with those obtained from the simplest model, residual gain analysis (RGA) . All correlations were above 0.93 and most were 0.98 or higher indicating that the school value added indicators would hardly be altered by the use of the more complex computations. Further analyses are presented in the technical reports and reinforce the conclusion that the use of the extra complexities did not make much difference to the indicators.

Prosser, R., Rasbash, J. and Goldstein, H. (1990) ML3, software for three-level analysis. Users guide for V2. Institute of Education, University of London.

There is also a module in SAS.

The more complex, less 'readily understandable' models can be seen to have made little difference for a first stage of information feedback to schools. The use of all types of models could be thoroughly explored in the second stage of data analysis.

Columns 9 and 10 show the percentage of schools that would be deemed to be, statistically, significantly above or below average in terms of the relative progress of their pupils (value-added)[7]. The more complex models sometimes led this to increase, sometimes to decrease, but essentially did not have a great impact. There should be no routine use of the arbitrary 0.05 significance level; substantive differences are the differences of concern. Unfortunately it is difficult to arrive at consensus on the size of differences that are considered to be of substantive significance.

Column 8 shows the value of rho, the proportion of total variance that might be attributed to the school. The additional complexity of the method of analysis had little impact. (Rho is described in the glossary)

Columns 6 and 7 show the percentage of variation that is 'explained' by the model for pupils and for schools. The striking feature here is not that the models have an impact but the differences between primary and secondary schools. This could be seen as due to the greater impact of secondary schools on achievement but it could also be due to the greater segregation by ability that is found in secondary schools. Any school-level analysis is influenced by the fact that pupils are not randomly assigned to schools and segregation effects can have powerful effects on the data (cf. the discussion of mathematics classes in Annex C).

Table 3.6 Effects of modelling compositional effects , curves, and variable slopes

[7] The use of statistical significance as a criterion is not entirely satisfactory

0	1	2	3	4	5	6	7	8	9	10	11
of sis	Dataset	Pupil prior ach.	Mean pupil prior ach.	curve fitted	pupil prior ach. slopes vary	Pupil var. explained	School var. explained	Intra-school correlation	% schools worse @ .01	% schools better @ .01	r with RGA
MARY DATA SETS											
	95 Trial Avon							0.25	16	12	
le VA	95 Trial Avon	•				45	22	0.32	17	18	0.99
	95 Trial Avon	•	•			46	27	0.31	17	17	0.98
	95 Trial Avon	•		•		46	20	0.33	16	18	0.99
	96 Trial							0.21	11	12	
le VA	96 Trial	•				55	60	0.22	12	13	0.94
	96 Trial	•	•			55	60	0.22	15	14	0.93
	96 Trial	•		•		56	60	0.22	12	17	0.95
ONDARY DATA SETS											
	.English 96							0.27	26	34	
le VA	English 96	•				52	83	0.12	20	22	0.99
	English 96	•	•			52	83	0.11	21	19	0.99
	English 96	•		•		53	83	0.12	21	22	0.98
	English 96	•			•	54	78	0.15	20	23	0.98
	maths '96							0.31	27	34	
le VA	maths '96	•				66	90	0.11	21	21	1
	maths '96	•	•			66	90	0.11	20	20	0.94
	maths '96	•		•		66	90	0.11	21	21	0.99
	maths '96	•			•	67	87	0.15	19	23	0.99

We turn now to summarising the findings for each of the additional features in the models.

Finding 8. The average level of prior attainment in a school sometimes had a significant effect on relative progress and sometimes did not.

A compositional variable is composed of an aggregation of pupil-level data to the group level. The group is often the school but could also be a department in a school or a class, ie a teaching group. An example of a compositional variable is the mean prior attainment for the whole cohort. The composition of the school in terms of the characteristics of its intake, seems highly likely to have an impact on how the school functions. However, the impact rarely appears to be strong statistically and it is sometimes present and sometimes absent. This is an area needing further work both qualitative and quantitative and there are problems of the correlation of this variable with the pupil-level data and whether or not the measure should be 'centred' for each school, i.e. measured from the school mean.

The composition of the intake is not an issue for internal school improvement, since the school has to work with whatever pupils it has, but it could be an issue in situations in which there was local control over the distribution of pupils among schools. Furthermore the composition of classes within the school may be important (streaming and setting practices) and could well be a topic explored in the second stage exploratory data analysis, but only if teaching group was recorded in the dataset or added to the national datasets when these were analysed locally.

The findings regarding school compositional effects, with respect to average attainment of the intake, in the various data sets are summarised in Table 3.7

Table 3.7 Compositional effects for prior attainment .

Primary 1 (PIPS)	The compositional effect based on a concurrent measure of developed ability 'explained' a further 11 % of the school variance, making important differences to the value-added scores
Primary 3 (Avon95)	A peculiar negative effect was found, probably due to problems with the end of key stage assessments in schools
Primary 4 (Various LEAs)	The odd negative compositional effect found in Primary 3 disappeared and a more usual, statistically significant, positive effect was found
Secondary 1	No significant compositional effect
Secondary 3 Pilot	Statistically significant compositional effects were found, of different magnitudes depending upon the subject, but not making large differences to the value-added indicators.

It can be seen from the summary that compositional effects were inconsistent The possibility that they are there and can be easily measured (requiring only aggregation of the data already in use) makes them a feature that would be routinely investigated in the stage two exploratory data analyses. The fact that they appeared in some subjects and not others cast doubt on the effect being attributable to 'the school'. In secondary schools the variance of the intake (ie the extent to which teachers were facing a large range of aptitude among pupils) was also tried as a compositional effect but did not contribute significantly.
These inconsistent results perhaps illustrate a warning issued by Aitkin and Longford (1986, p.13) in their classic discussion of models :

> ..... *it will be found quite generally that the standard errors of individual-level variables aggregated to the school level are very large, when the individual-level variables are also included.*

Use of means in the models makes the results subject to the instability that is inherent in a mean-on-means analysis and is illustrated in the glossary under the heading 'correlation, the slippery statistic'.

Finding 9. A curve provided a slightly better fit to the data than a straight line, but had little effect on value-added indicators

In most datasets a curve fitted very slightly better than a straight line but for the vast majority of pupils this made no substantial difference. Over the major part of the actual range in the data the linear and curvilinear predictions were highly similar. Thus it was not surprising to find the value-added indicators almost indistinguishable, correlating 0.99 and 0.98

in the secondary datasets for example. It must also be noted that the interpretation of curves must consider errors in the predictor as well as the outcome. Any ceiling or floor effects in either can have an effect on the apparent relationship. Schools with exceptional intakes (very low or very high) could possibly benefit in terms of accuracy from having their data analysed both with lines and curves in the second stage of analysis, but for the overwhelming majority the data showed lines to be adequate.

Finding 10 Whether or not slopes were modelled as varying had little or no impact on value-added indicators (ie there was no evidence that schools were differentially effective with pupils of different prior attainment levels.)

It can be noted immediately , from column 11 in Table 3.6, that the effect of modelling 'slopes vary' had no impact, with correlations of 0.99 and 0.98 in English for secondary schools.

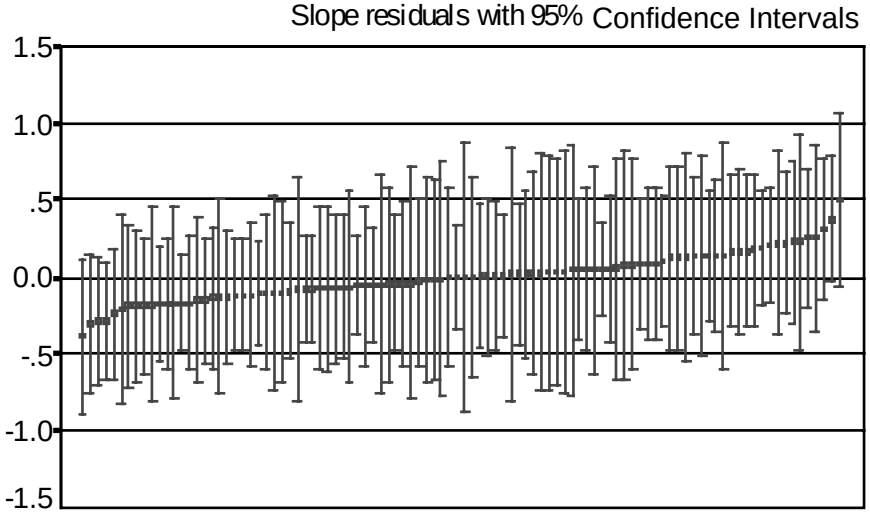
The concept of 'varying slopes' (differential effectiveness for pupils with different prior attainment levels) is explained in the glossary but need not be of concern. It is an area in need of substantial research, providing a topic for stage 2 investigations. Schools must not be held responsible for merely chance findings,

Slopes are slippery statistics, easily changed by the results from one or two pupils, as explained in the glossary. Thus it was not surprising to find that very few varying slopes in the datasets could be considered 'significantly' different from the average slope, applying the usual .05 criterion. This finding is illustrated by Figure 3.1 reproduced from the Primary technical Report No. 4. For a selection of schools it shows the extent to which slopes differed from average as the centre of the vertical lines. The vertical lines ('confidence intervals') suggested that these differences could well be anywhere in the range shown from top to bottom of the vertical line. Since the range included zero for every school, not one of the slopes could be considered reliably different from an average slope.

Basically, varying slopes by intake could be ignored in the primary datasets.

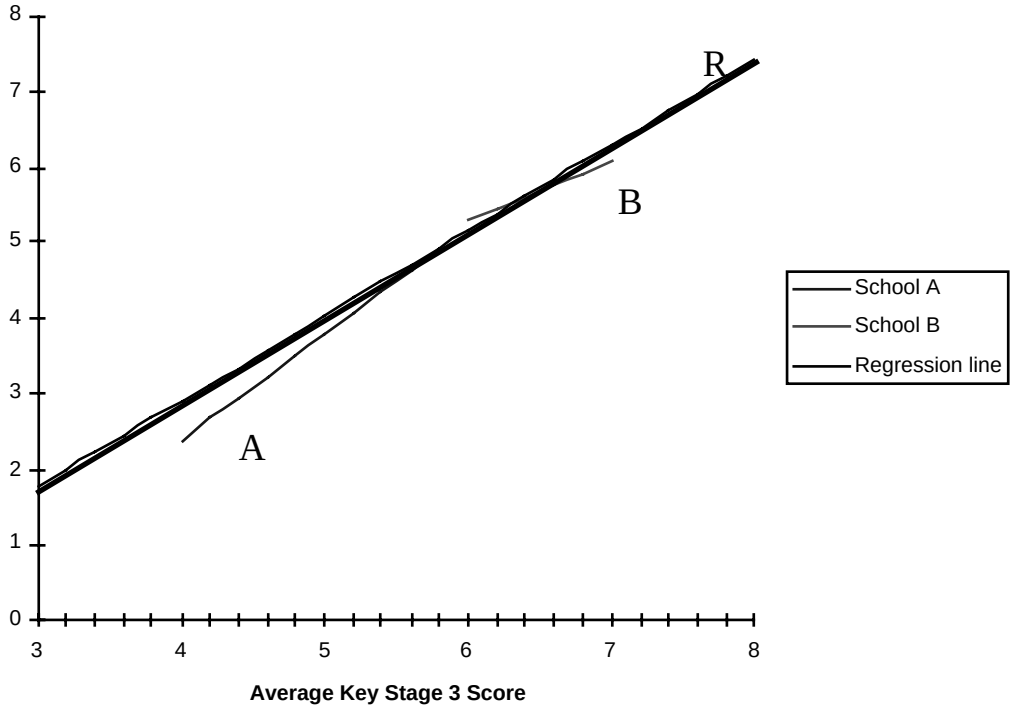
In secondary schools the slopes are more likely to show statistically significant differences because of the larger numbers of pupils. However, the graph in Figure 3.2 shows the two most extreme slopes from the 1996 secondary pilot. *Over the range of their actual data* the slopes hardly departed from the overall regression line . It they were extrapolated beyond their range, then the discrepancies would be considerable. This situation causes confusion in the interpretation of differential slopes. Its impact is seen in the lack of difference on actual school value added scores when slopes are or are not considered to vary within schools: correlations of 0.99. Until differential slopes have been seen consistently over several years and can be related to actual teaching or organisational practices, any interpretation must be extremely cautious in order not to waste people's time and effort with false alarms (what statisticians call Type I errors).

Figure 3.1 Uncertainty about slopes:



From the Primary Technical Report no.4

Figure 3.2 Statistically significantly different slopes from the secondary pilot



3.3 Question C. Are there benefits to be gained from the inclusion of additional variables (a) for internal school management?....(b) for publicly available information?

The models discussed above could all be investigated without the need to store any further information beyond the test and examination data linked

to schools and pupils. In the following paragraphs the addition of other kinds of information is considered.
The additional variables considered are: social indicators, sex, mobility measures, attendance data and examination syllabuses/boards.

Finding 11. In the majority of schools, social indicators contributed little to the prediction of relative progress (value-added) but the presence of atypical schools requires that attention is paid to this variable.

Social indicators measured for individual pupils (ie *pupil-level* social indicators) generally correlate with achievement only 0.3, thus accounting for only 9 percent of the variance whereas prior achievement generally accounts for about 50 percent. Pupils' potential cannot be accurately judged by their home background. This general finding remained true. The achievements of individual pupils are not well predicted from knowledge of their home background.
However, the impact of *school level* measures, such as the percent of pupils in each school that were eligible for free school meals, was found to be substantial in some datasets acquired in the second phase of the project. Some 10 percent of primary schools in the sample had over 60 percent of pupils taking free school meals. These were schools with smaller than average enrolments (about 22 pupils as opposed to 33) and their relative progress (value-added measures) was about 0.2 of a level lower than the average school that had 20 percent of pupils taking up free school meals. In the secondary dataset there were about 6 percent of schools with over 60 percent of pupils entitled to free school meals. Whilst the overall impact was weak, with these extreme scores the effects on individual schools would need to be considered. It is precisely this kind of attention to outliers that is important in the proposed exploratory data analysis in the second stage.
However, whether the percentage of pupils eligible or entitled to free school meals should be included in value-added indicators is not a matter that can be decided by statistically significant findings, nor even findings of substantial differences. The inclusion of free school meals in the predictions could be seen as serving as an excuse, saying that less would be expected of pupils in schools with a high percentage of pupils eligible for free school meals. Given the same prior attainments, should not the same subsequent attainment be expected? The critical issue for education is to seek ways to ensure that this equivalent progress is made.

Finding 12. The sex effect was in favour of girls in some subjects

The impact of including sex in the models used to produce value-added measures was slight if the outcome was a total score, but even slight differences can become important when a school has an atypical concentration of pupils, such as in single sex schools.
The consideration of value-added differences between boys and girls is discussed in Section 4 on using value-added information. If there is to be

publication of value-added indicators on a whole school basis then it would seem to be important to have comparisons that take single-sex schools into account. As with all the additional variables, their influence is small as long as the intake to the school is not severely atypical. The whole school indicators that might be used are displayed in Figure 3.3 for single-sex and mixed schools. Whilst the advantage to girls is small in any one subject (except English), the girls' schools do occur on the upper side of the cluster of means. However, this was at GCSE. It is generally found that the pattern reverses at A-level.

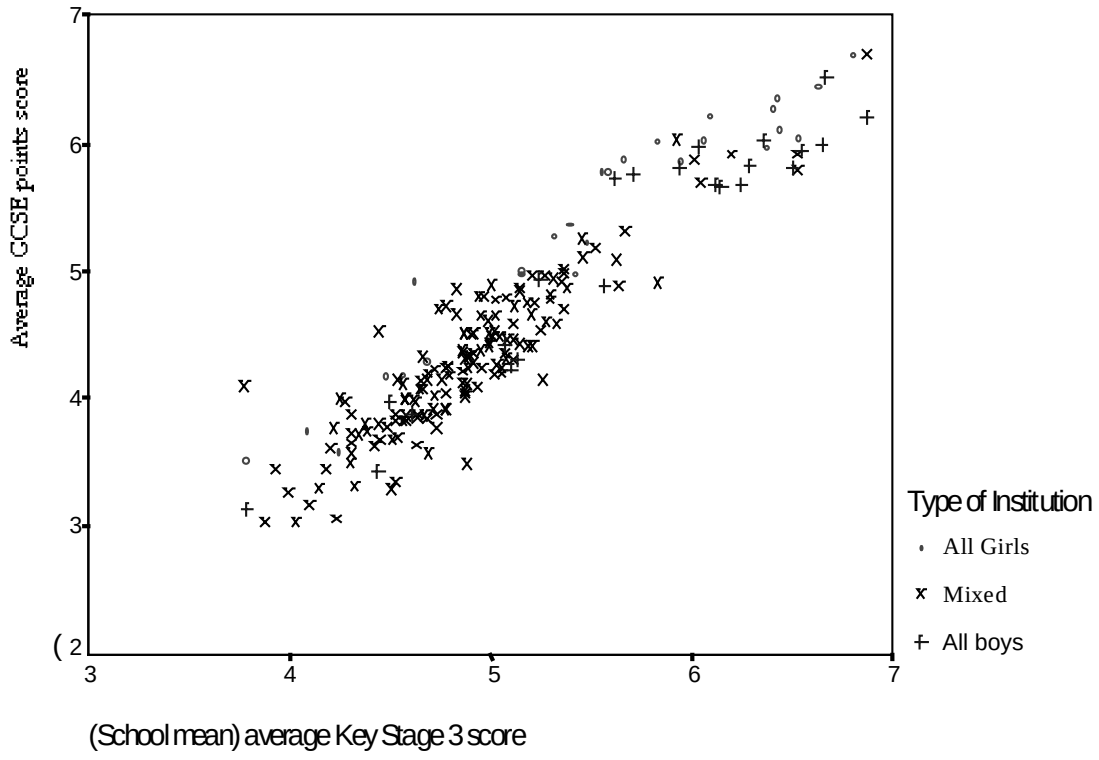
	Effect on Relative progress (value-added)
Primary 1 (PIPS)	Sex statistically significant in MLM but had no impact on school value-added measures.
Primary 2 (Avon95)	Sex not statistically significant and slightly in favour of boys
Primary 4 (Various LEAs)	Girls made 0.2 levels more progress in English but boys made 0.17 levels more progress in mathematics and 0.10 in science. (p.12)
Secondary 1	Sex statistically significant and about one third of a grade on average in favour of girls, but the same for all schools
Secondary 2 Pilot	Again a statistically significant effect of about one third of a grade in favour of girls. In English, girls were making two thirds of a grade more progress on average than boys. In mathematics, boys were making about a tenth of a grade more progress on average than girls.

As with free school meals, there is the issue that including the sex variable in the modelling serves as a kind of excuse. Again the educational issues are the extent to which any schools overcome the differences in average attainment between the sexes for this age of pupil on the kinds of examinations in which this difference occurs and whether the observed differences in a school are similar to those in other schools or substantially different.

Finding 13. Mobility and attendance presented substantial and unresolved issues

Schools cannot 'add value' to pupils who are not present and it appears to be the case that, for some schools, the pupils who have been present throughout the key stage are a minority of pupils on the roll. The problem of pupil mobility can be managed in the school's own interpretation of its own data but would present severe problems for fair performance indicators relating to value-added if these were to be public.
The analyses and pilot studies were able to cast some light on the magnitude of the problem and to address an important question regarding the characteristics of mobile pupils. If mobile pupils were simply a random sample with characteristics similar to those of less mobile pupils, then whether or not they were included in value-added indicators would be unimportant. If, however, they were not typical, the issue of which pupils would 'count' would become important.

Figure 3.3 School means showing single sex and co-educational schools.



A pupil was defined as mobile if he or she had changed school in other than a routine progression. Obtaining such data is not simple unless the local patterns are known with regard to nursery, infant schools, junior and middle schools, for example. Thus mobility of pupils in an area is probably best assessed in the local area, a possible contribution from LEAs. The existence of a unique pupil ID would clearly be of assistance.

Table 3.8 Estimated mobility figures for pupils in the primary pilot

Mobility	Number of pupils	Percent of pupils	Average value-added score at KS2
No change of school	3954	48	0.06
Infant changed to junior	2855	34	-0.06
Transient pupils	1415	17.2	-0.04

About 17 percent of pupils in the Avon datasets had moved schools more than necessary but this mobility was concentrated in a minority of schools. Because records followed the pupils in Avon, there were value-added measures available for these mobile pupils. The differences between mobile and non-mobile pupils in value-added were 0.1 of the average KS2 level in 1995 and 0.07 in 1996, both differences being in favour of the non-mobile pupils. These were small differences but could have an influence in those schools with high proportions of mobile pupils.

At the secondary level, over the shorter time span of two years, there was evidence of *attrition* (pupils who had been present with an end of Key Stage 3 score but were not represented in the GCSE results) and *in-flow*, pupils joining the school and taking GCSEs but for whom end of Key

Stage 3 data were not available. Table 3.9 shows that the pupils who left before GCSE results were from the lower end of the attainment range in greater proportions than from the higher end. Furthermore schools reported attrition rates of up to 20 percent.
On average, those who left had an end of Key Stage 3 score of 4.2, which was 0.8 of a level lower than pupils who remained.

Table 3.9 Characteristics associated with mobility in Key Stage 3

Average Key Stage 3	Retained (28,038)	Left (2168)
below level 4	87%	13%
level 4-7	96%	4%
above level 7	98%	2%

One serious problem is knowing whether or not pupils who leave at this stage of schooling go to another school or leave education early, before the end of 'compulsory' schooling. Pupils who appeared to have joined the schools during the two year GCSE period, the 'in-flow', had substantially different GCSE results. They had been entered for about 6 rather than the typical 9 GCSEs. However, they had average GCSE scores of only 0.4 points per examination entry lower than non-in-flow pupils.

Table 3.10 Differences between pupils in the school throughout the Key Stage and pupils who arrived at the school during the two years before GCSE.

	Pupils on 1994 CRF (20,379)	New pupils (1096)	Effect Size[8]
Average GCSE score (SD)	4.6 (1.52)	4.2 (1.63)	0.26
Total GCSE score (SD)	41.2 (17.24)	27.4 (19.98)	0.80
Number of GCSEs (SD)	8.7 (1.18)	5.9 (5.91)	2.37

The percentage of "new", in-flow pupils for each school ranged up to 20%.
In summary, the value-added measures could not be considered an entirely *valid* representation of the academic work of the school if large numbers of pupils had not been in the school throughout the relevant period of time. The extent to which this was the case varied considerably from school to school. Ways in which this problem might be accommodated in a value-added system are discussed in Section 5.
The essential lessons for a value-added system are that records must accompany pupils as they move schools, movement between schools should be documented and the regular recording of the rate of attendance at each school might also be needed. Further research on the impact of mobility on pupils is desirable, especially as the National Curriculum was introduced partly to enable pupils who were mobile to have a better chance to maintain their academic progress. Evaluation of this impact is needed. To learn from the data it will be essential that the data are better validated

[8] The standardised mean difference between two distributions. In this case the Effect Size was the difference between the mean for non-changing pupils and that for 'new' pupils, divided by the pooled standard deviation.

than current truancy data, particularly as any public reporting of value-added would put pressure on such data, a topic touched on in Section 6. SCAA's decision to introduce optional tests for use in Year 4 could make a two year value-added system more feasible but the problems of mobility and small sample sizes would still need attention.

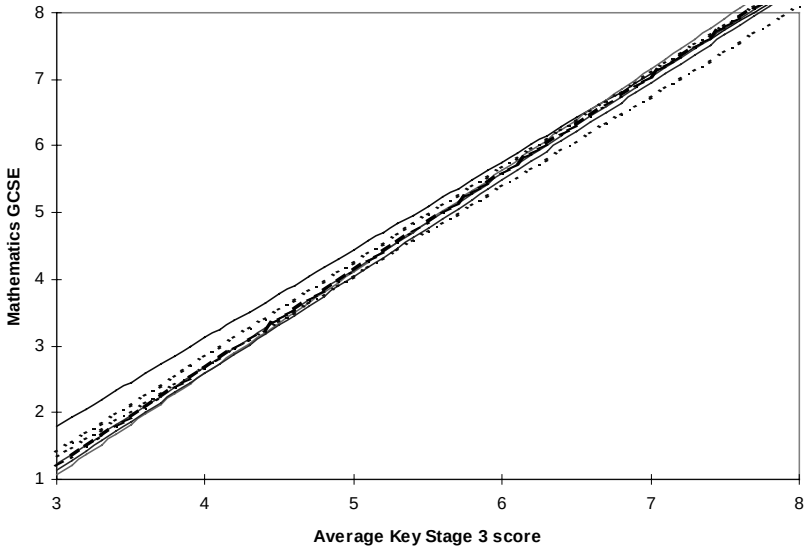
Finding 14. There were differences between syllabuses and examination boards in value-added.

With a single national curriculum for the end of key stage tests, the issue of syllabus comparability does not arise in primary schools or at Key Stage 3. In secondary schools, the issue of the comparability of difficulties[9] between examination boards is an important one. In both primary and secondary schools there will be variations in the apparent difficulties of subjects and these need to be monitored and reported regularly.

Staff in secondary schools often worry about the comparability of different examination boards. Is there an easy Board? The answer is no, but there are differences from year to year and from syllabus to syllabus and these will need to be made public or in some way taken into account in a value-added system in order for it to be perceived as fair. Whilst the examination boards have been sufficiently comparable for candidates to have few concerns, the situation changes as schools are increasingly evaluated by examination data.

The situation with respect to Mathematics in 1996 is illustrated in Figure 3.4

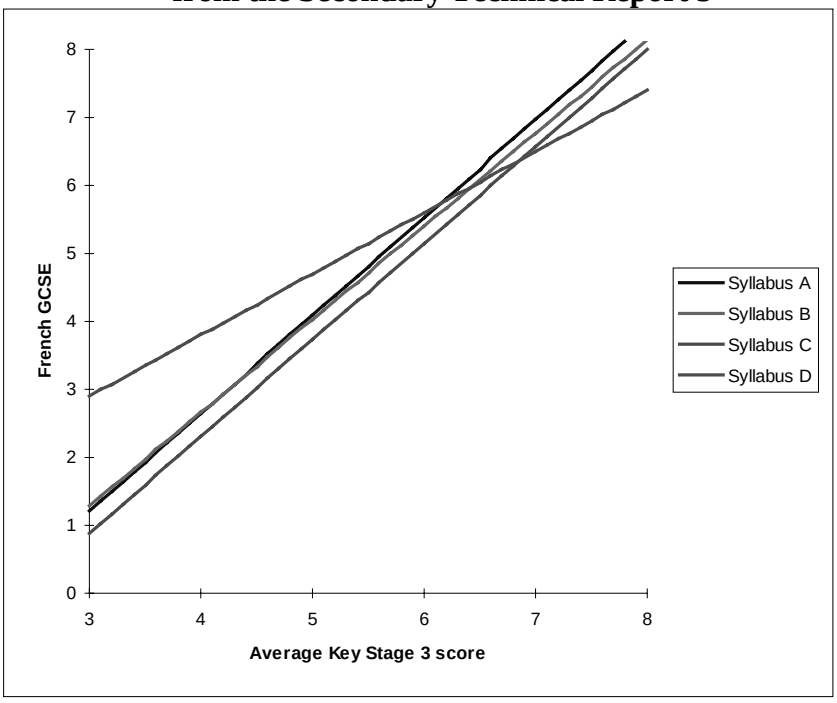
Figure 3.4. Regression segments for Mathematics GCSE results, 1996: from the Secondary Technical Report 3



[9] 'A subject or syllabus is defined here as 'difficult' if the grades or levels are generally lower in that subject than in other subjects for pupils with the same prior attainment. This could be due to severity of grading. 'Difficulty' is used as a shorthand.

The regression lines in Figure 3.4 show good consistency across the syllabuses in Mathematics. However, this was not the case for all subjects, as shown in Figure 3.5.

Figure 3.5. Regression segments for French GCSE results, 1996: from the Secondary Technical Report 3



Notes: Data for Figure 3.5

Syllabus	Mean (Average Key Stage 3 score)	Mean (GCSE grade)	Mean (Residual)	Number of candidates	Number of Schools
A	5.4	4.7	0.00	8710	87
B	5.5	4.7	-0.12	899	13
C	5.3	4.2	-0.37	3594	42
D	4.5	4.3	0.73	2025	34

Over 95% of candidates for syllabus D had an average Key Stage 3 score of 6 or less

One syllabus apparently served the purpose of providing a French course in which lower achieving pupils could achieve success. This could be taken as an indication of 'suitability for purpose' and is not necessarily a situation that needed to be changed. It did need to be reported. A national value added system should obtain the regression lines for each syllabus and report them, including the 'correction factors' as reported in the Dearing *Review of 16-19 Qualifications.* In this way choice and diversity need not be impeded yet lack of comparability becomes public knowledge. To attempt to make all subjects equally 'difficult' would mean that some subjects would often have no failures because of the high prior attainment range of the intake, whilst other subjects would be able to award almost no high grades because of the low prior attainment range of the intake.
Figure 3.4 shows that the different syllabuses for mathematics were quite closely aligned in 1996. However, the small differences could affect the value-added scores since a school generally enters candidates for only one syllabus. Syllabus and school are confounded.

3.4 Conclusions, related to the initial questions

Question A. Is there a basis for a statistically valid and readily understood national value-added system in the near future for use in internal school management?

Yes.

Since there are adequate predictors and outcomes in place for the key stages investigated in this project, the elements of an immediately workable value added system are already in place for Key Stage 2 and Key Stage 4. This finding probably generalises to the other key stages but only if each measurement used for value-added purposes has the consistency and reliability that is most effectively attained by external marking.

For purposes of internal school management, there is no reason to delay the introduction of a national value-added system based on the first seven findings in this section.

For purposes of accountability involving publication of school by school value-added indicators, there are several complications:

- value-added indicators show low stability from year to year. The extent to which this variation is under the control of schools is not known but certainly one year's data would be potentially misleading for many schools. Delay of publication until three years' data are available, and publication in the form of three year averages, would seem to be desirable.
- special attention has to be paid to 'outliers', the small proportion of schools with highly atypical intakes, such as single sex schools and schools with high proportions of free school meals uptake. Their value-added scores might best be published both with and without an adjustment for their special circumstances.
- any indicator should be based on the work of all pupils in the school, counting each pupil equally. With high rates of mobility among some pupils, exactly which pupils should or should not be counted is a difficult issue, as is the treatment of data from pupils with erratic or low attendance. Furthermore the concentration of high mobility pupils in some schools requires special consideration. Methods of recording mobility and attendance rates need to be uniform and objective in order to be credible.

In order for a value-added system to work smoothly and cost-effectively there needs to be an infrastructure developed as quickly as possible. Desirable immediately are a unique pupil identification number for each pupil and the capacity of schools to store, retrieve and pass on, by electronic means, the following information:

- the unique pupil ID;
- An enrolment record that is updated whenever a pupil changes school, whether in the course of normal progression or due to mobility;
- For each pupil, the prior attainment details (average level as a minimum) linked to the school/centre at which the measures were obtained (This cross-checks enrolment data.);
- The percentage of the key stage for which the pupil attended the school (a measure covering both mobility and attendance).

Given the considerable work associated with recording attendance and linking this consistently in all schools with value-added data, alternatives may have to be considered and some are suggested in Section 5.

Question B. Would a second stage of data analysis be necessary prior to making value-added indicators public on a school by school basis? (Findings 8 through 10)

Yes, but not of such magnitude or clarity as to justify large expenditure against the education budget.

Question C. Are there benefits to be gained from the inclusion of additional variables such as sex, social indicators, mobility measures, attendance data and examination syllabuses / boards...

(a) for internal school management?

Schools will have access to such data on their own pupils, and to even more important and relevant data in terms of *how pupils were taught*. It will be important that schools are not given over-generalised statements about insecure findings. This could needlessly distract them from their own well-informed insights into the work of their own school. On the other hand, some schools may have patterns in their data that are meaningful and useful and would be overlooked without a second stage of analysis.

The extent to which the second stage with additional variables informs or confuses will have to be a matter resolved by several empirical trials. The conclusion is not foregone, in either direction.

(b) for publicly available information?

Adjustments to the simple analyses are likely to be needed only for highly atypical schools. However, before information that could impact on a school's reputation is made public, the findings need to be as secure and fair as possible. The method of analysis is only part of the approach and often makes little difference. More important differences are made by taking account of additional information, ie the inclusion of additional variables in the second stage of analysis. The most influential variables to include would seem to be sex and possibly mobility. Language effects could be important in some schools. There are unresolved tensions between taking factors into account and not lowering expectations or making excuses. Only when it is known from experimentation what differences schools can make, will accountability be reasonably fair.

This section concludes with a 'finding' that arises not from analysing the data but from the attempts to collect data.

Finding 15. The need for Electronic Data Interchange (EDI)

A major source of time-consuming problems with most data analyses is the setting up of the files. Research institutions have been used to 'cleaning' data, a procedure that often involves simply discarding cases with missing values or using some statistical method to 'impute' the likely value of missing information. In dealing with data for a pupil-by-pupil value added system such techniques are not acceptable as each case will be scrutinised by schools and must be accurate. Unless the datasets are transferred electronically this one hundred per cent accuracy is extremely difficult.

Even with EDI problems will arise with such problems as disk formats and accessible files. Without EDI there are likely to be numerous instances of difficulty in matching the data and in dealing with transcription errors. To quote the secondary pilot report:

> *'The quality and precision of the input data varied widely from high quality information retained electronically on schools' management systems, through the bulky but data rich Class Record Forms, to reams of individual student record slips.'*

The processing of such diverse inputs is time-consuming and uneconomical. Furthermore, it should be noted that many schools were not able to participate in the pilots because they had not retained data. The pilot studies, along with years of experience in the CEM Centre in providing schools with analyses, makes us strongly recommend that any national system must be based on electronic data interchange using unique pupil identifiers.

3.5 Summary

Sections 2 and 3 have summarised the findings from project activities over a 22 month period. The conclusions are very positive towards a national system. A clear, simple, readily understandable first stage is viable and there are sufficient complications to justify a second stage of exploratory data analysis. The difficulties likely to arise from problems in data storage and handling could be overcome by the use of electronic data interchange using unique pupil identifiers. A problem that varies in its intensity from school to school is that of attendance and mobility. Secure methods of recording these additional variables need to be developed.

Section 4 considers the type of feedback to schools that could be immediately implemented and Section 5 then takes up the issue of how a national value-added system might be organised.

SECTION 4: USING VALUE-ADDED DATA

Some of the ways in which value-added data might be used are considered here, first within schools, for internal management purposes, and secondly for the purposes of public accountability, possibly in School Performance Tables. Readers already familiar with one of the value-added systems already in widespread use may prefer to turn directly to section 4.7.

4.1 Information on each pupil's relative progress

Teachers are interested in the performance of each and every pupil. A tabulation such as that shown in Table 4.1 (Pupils' value-added scores) presents the data in a familiar format. The real data presented in Table 4.1 show progress from Key Stage 1 ("KS1 average level") to Key Stage 2, using the average level across three required subjects (English, mathematics, and science). This illustration is from the primary sector but equivalent procedures apply for secondary schools.

The predicted level can be taken simply as the level the pupil would have received if he or she had made average progress, based on the pattern in the whole *national* dataset *that year*. Interest focuses on the differences between statistically predicted score and actually attained score.. The list can be ordered according to these differences or alphabetically. Table 4.1 is ordered so that the pupil who made the most progress is at the top of the list, showing progress of 0.64 of a level greater than predicted. This is the 'value-added', a measure of progress relative to prediction.

This listing of differences (i.e. of value-added scores for each pupil) can be followed by discussions of each pupil with the headteacher or an adviser in primary schools or with the Head of Department in secondary schools. By keeping an individual teacher's data confidential in the first year or two of participation in a value-added system, the anxieties that surround this approach can be alleviated. The experience of many schools is that open discussion of the data becomes more acceptable after some years of participation, a point illustrated by one response to the question 'How would you advise other headteachers to use value-added results?':

> *Very slowly! It takes about 2 to 3 years for staff to get to grips with it. Don't give out too much data to start with, and once you get into it don't over-interpret it. You need a senior member of staff to be very au fait with it.*
>
> Value-added National Project, Secondary Pilot

Table 4.1 Pupils' value-added scores: School A

1	2	3	4	5
SEX	KS1 average score (level)	KS2 average score (level)	KS2 predicted score (level)	Value-added Difference: column 3 minus column 4
**M	2.59	5.00	4.36	0.64
M	2.32	4.67	4.06	0.61
F	1.2	3.33	2.85	0.48
F	1.53	3.67	3.21	0.46
M	1.58	3.67	3.26	0.41
F	1.71	3.67	3.40	0.27
M	2.04	4.00	3.76	0.24
M	1.75	3.67	3.44	0.23
F	1.49	3.33	3.17	0.17
M	2.44	4.33	4.19	0.14
M	1.53	3.33	3.21	0.12
F	1.88	3.67	3.59	0.08
F	2.22	4.00	3.95	0.05
F	1.33	3.00	2.99	0.01
M	1.33	3.00	2.99	0.01
M	2.00	3.67	3.72	-0.05
M	2.00	3.67	3.72	-0.05
F	2.04	3.67	3.76	-0.09
M	2.04	3.67	3.76	-0.09
M	1.75	3.33	3.44	-0.11
M	2.48	4.00	4.23	-0.23
M	1.59	3.00	3.27	-0.27
M	1.92	3.33	3.63	-0.30
M	2.04	3.33	3.76	-0.42
M	1.43	2.67	3.10	-0.43
M	1.75	3.00	3.44	-0.44
M	1.52	2.67	3.20	-0.53
F	2.15	3.33	3.88	-0.55
F	1.08	2.00	2.72	-0.72
M	1.76	2.67	3.46	-0.79
F	2.1	3.00	3.82	-0.82
*M	0.92	1.67	2.55	-0.89
F	1.28	2.00	2.93	-0.93
F	1.04	1.67	2.68	-1.01

*Referred to later a pupil J.

**Referred to later as pupil K

The discussion with teachers will include a search for the possible reasons for the relative lack of progress of some pupils (e.g. those at the bottom of the list in Table 4.1) with a view to identifying any problems that can be avoided in subsequent years. Equally, the particularly good progress of other pupils (those at the top of Table 4.1) will be discussed to see if

lessons can be learnt as to how to promote such good progress in the future among other pupils.

In addition to the concern for each individual pupil, the school will be interested in the general level of performance. In general, in a school in which pupils are generally making average progress, about half the pupils will have value-added scores that are positive and about half will have value-added scores that are negative.

At the classroom level the samples will generally contain fewer than 35 pupils and the margins of uncertainty on the value-added score will therefore be large. Although small samples are seen as a problem statistically, for internal school use this is not necessarily the case. Indeed, small numbers of pupils can make the analysis easier and more transparent. The variability inherent in small samples is readily apparent when teachers examine their own data. For example, teachers intuitively conduct sensitivity analyses, recalculating what the value-added would be if they discounted some pupils, or analysing outliers noting, for example, a pupil whose long absences provided a reason for not counting that pupil into the value-added system.

In the context of internal school use, the issues of attendance, motivation, mobility, friendship groups, problems at home, effects of part-time employment and other responsibilities that some pupils carry, might all be discussed. However, most of these considerations relate to aspects that are difficult to alter or are, perhaps, the personal concerns of the pupil and either not known to staff or not appropriate for general discussion. The most important discussions will focus on teaching and learning activities: the **alterable variables** (Bloom, 1979). Are there changes that can be made that lead to better outcomes?

Scatter plots

The table alone could provide a sufficient basis for discussion of value-added results but some staff, governors or advisers might find the scatter graphs a useful visual representation of the same data. The data in table 4.1 is graphed in Figure 4.1 below.

There will be some temptation to draw a ‘school line’ for comparison with the national line but this may not be advisable. The school line will be subject to the influence of outlying scores and could vary substantially according to which pupils were counted. In other words, the school line, based on small numbers of pupils, will be subject to large margins of uncertainty.

Figure 4.1.(a) The plot of each pupil’s KS1 average score and KS2 average score

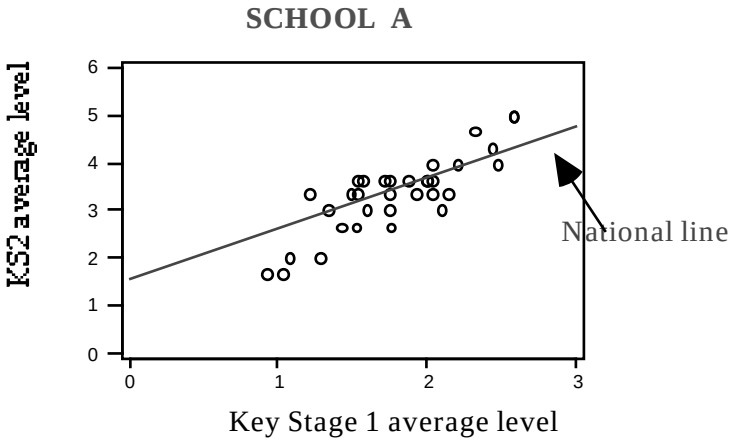
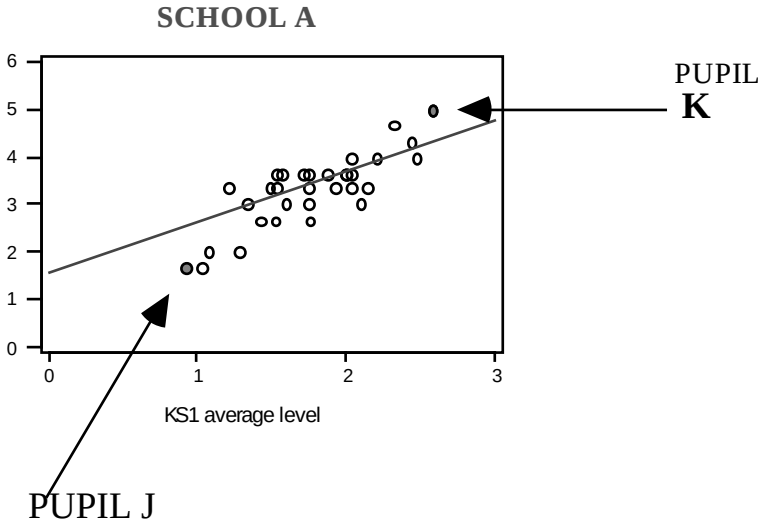


Figure 4.1.(b) The same plot with pupils J and K identified (compare with Table 4.1).



The sloping line on the graph is drawn from the national data and must be provided to the school. (It can be derived from Table 4.1 by plotting predicted scores but is best provided directly to avoid uncertainty). The line can also be used to indicate the likely progress of pupils in the following years, if the system did not change substantially. The equation of the line is useful to those schools with someone interested in using spreadsheets or statistical packages. In the data shown for the years 1992 to 1996, the equation of the line was:

(average KS2 score) = 1.55 + 1.08*(average KS1 score).[10]

This indicated a gain of approximately one and a half (1.55) levels across these years.

Clearly the minimum essential feedback is the Table, from which the line could be computed. Indeed the Pupil value-added score table could be provided on disk with an optional embedded program to plot the graph at the click of a button. Schools could use the graph or not, as they chose.

---

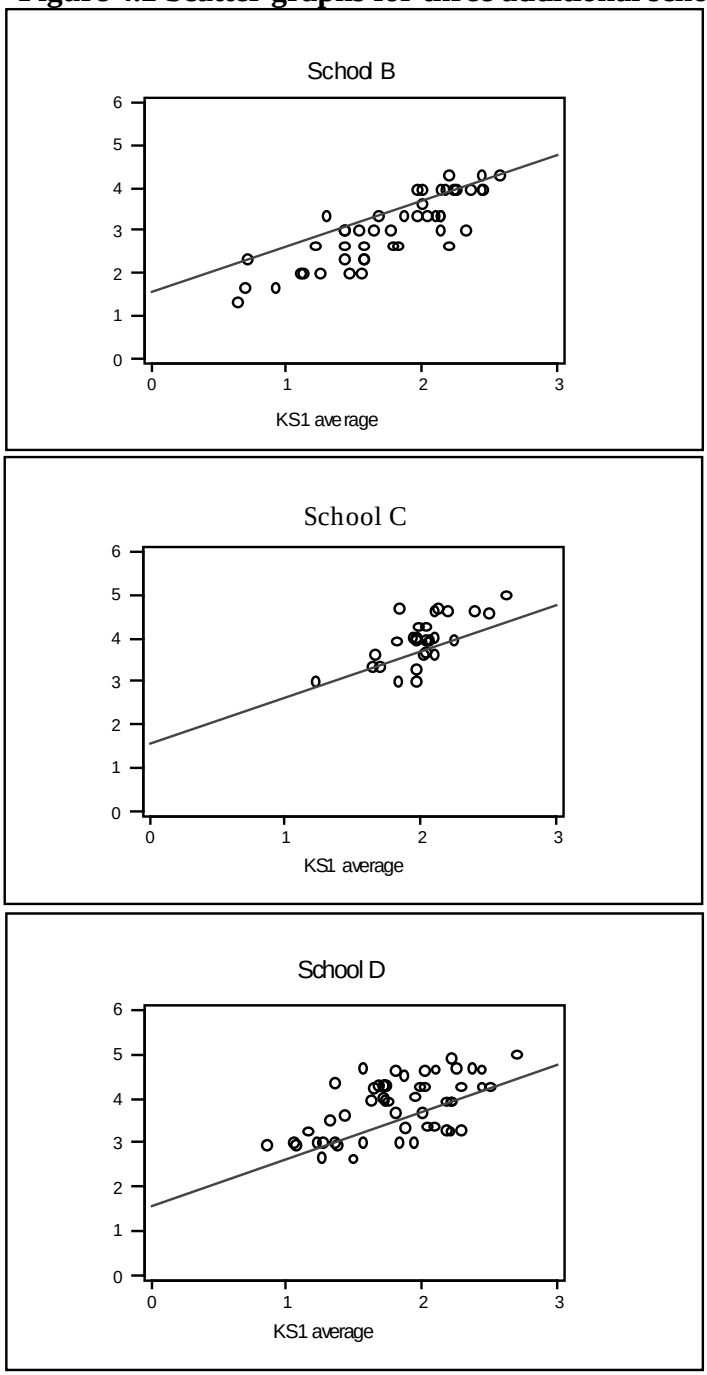
[10] To compute the predicted score for any pupil, that pupil's average KS1 score should be multiplied by 1.08. Add 1.55 to the result and the new result is then the average KS2 score, as predicted from the equation which was based on the general trend in the data in 1996.

Some schools would also wish to import the data into other packages, matching it with their own data for further analyses.

Because the eye can see patterns on graphs far more quickly than in tables of numbers, the graphs become particularly useful for scanning several sets of results such as those shown in Figure 4.2 these all use the same national line but the pupils differed and their performance differed. School B had a large proportion of pupils making less than average progress (points below the line), schools C and D both had many pupils making better than average progress (points above the line) and school C had a high attaining intake as shown by the points being largely to the right.

In a school, scatter graphs should be produced for each subject and this ‘at a glance’ quality is then highly valuable to headteachers.

Figure 4.2 Scatter graphs for three additional schools



The acceptability of the scatter plot as provided by Value-added systems such as ALIS, YELLIS and PIPS was endorsed by another provider, of the QUASE[11] system in the meeting of statisticians to discuss the Value-added National Project in which he described the scatter plots as

> ..... *the diagrams that the schools find most illuminating and useful in the whole report. The ones they tend to turn to are those kind of scatter plots with an overall line and then their data linked up.*

4.2 Feedback for the headteacher or Senior Management Team

It is increasingly common for the Value-added data to be used by senior management for a discussion with heads of department (secondary) or

[11]Quantitative School Self Evaluation, Dr. Ian Shagen, National Foundation for Educational Research

teachers (primary) upon the receipt of examination or test results. Additionally there will need to be a quick check on all subjects/ departments to see that they are obtaining reasonable results. Figures 4.3 and 4.4 show two ways in which the performance in different subjects can be summarised in easy-to-read displays. Each display includes an indication of a margin of uncertainty.

Figure 4.3 Paired Bar graphs of Value-added average scores for each subject, with 'margin of uncertainty' on the prediction represented by a short line

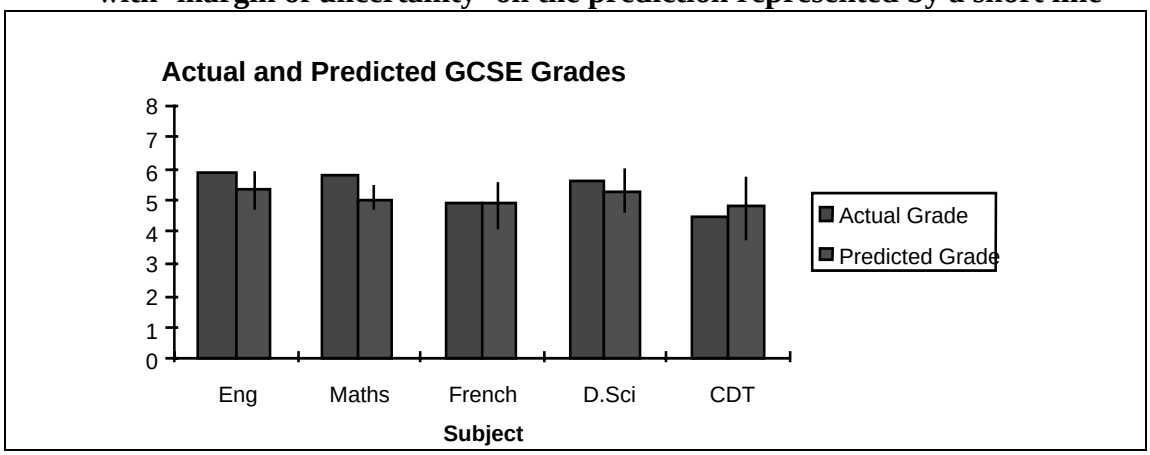
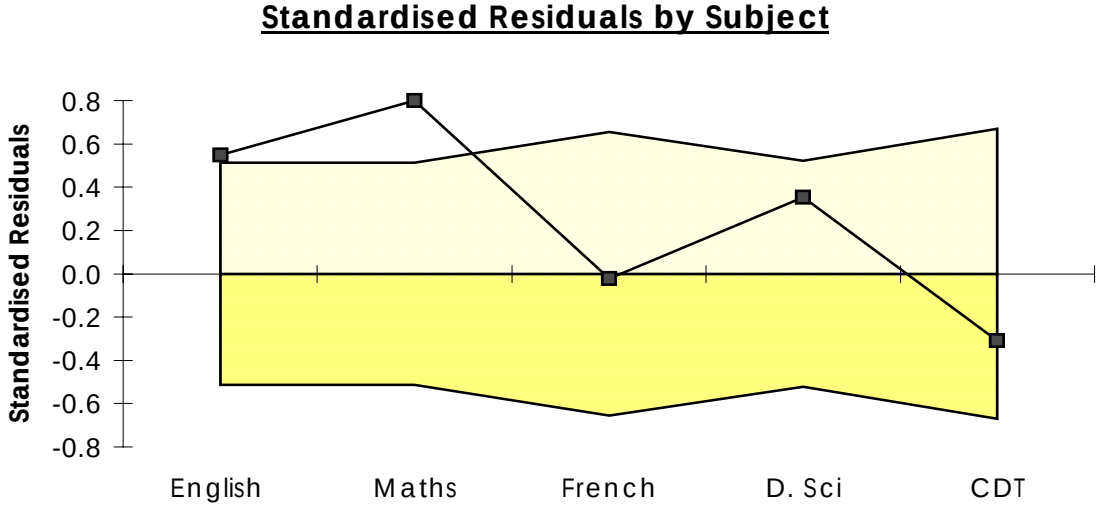


Figure 4.4 Statistical Process Control chart of the same data as in Figure 4.3



Figures 4.3 and 4.4 represent the same data in slightly different ways. Both show English and mathematics with better than predicted progress. Double science also showed better than average progress but only within the normal variation to be expected from year to year with different samples. The pupils in the French department had on average made average progress, hence their zero value-added score ie no difference from predicted. The CDT department had lower than predicted results but well within the expected variation and not, therefore, of particular concern for that single year.

Figures 4.3 and 4.4 employ the usual arbitrary choice that characterises statistical 'significance' testing: the use of the 5 percent criterion in drawing up the margin of uncertainty. Choice of the 5 percent level would mean that if the samples were simply random and there were no 'real' differences between schools, only 5 percent would be falsely identified as out of the normal range ('statistically significant'). However, as was reported in section 3, application of this significance testing tended to yield not 5 percent but 8 to 20 percent of schools out-of-range which suggested the schools did not look like simple random samples; they differed among themselves. But what should be considered an important difference? This is a much more difficult question to answer and, ultimately, the answer will depend upon showing that there are interventions that can reliably alter schools. Differences that can be removed are important; differences that no one has been able to alter have to be accommodated. There is not yet a body of educational research that can answer these fundamental questions about what is alterable, although the development of meta analysis is stimulating progress towards that goal.

Light monitoring

The display of the predicted and actual grades of a department enables a quick check to be made of the performance of pupils in each department. Sometimes data will fail to confirm but more usually it confirms and firms up patterns already recognised.

This light monitoring may be all that is needed: just a check that pupils in all departments have achieved, on average, reasonably in line with reasonable predictions. Glass (1979) wrote of 'fire-fighting'.

In addition to these summaries of the value-added data for each subject, headteachers will want to look at the value-added score of each pupil, across the subjects. For example, severe under-performance of a pupil in one subject may be interpreted differently if that pupil under-performed in all subjects rather than in just the one subject.

4.3 Tracking, monitoring and predicting

Many schools use last year's regression lines to make reasonable estimates of expected performance, pupil by pupil, or group by group. This is highly reasonable but there are a number of dangers. Teachers will use their professional judgement as to the emphasis to place on examination results with different pupils. For some pupils, constant pressure could be counter-productive for various reasons. This is the kind of judgement teachers are used to making in a profession in which day to day interactions are not routinised but human.

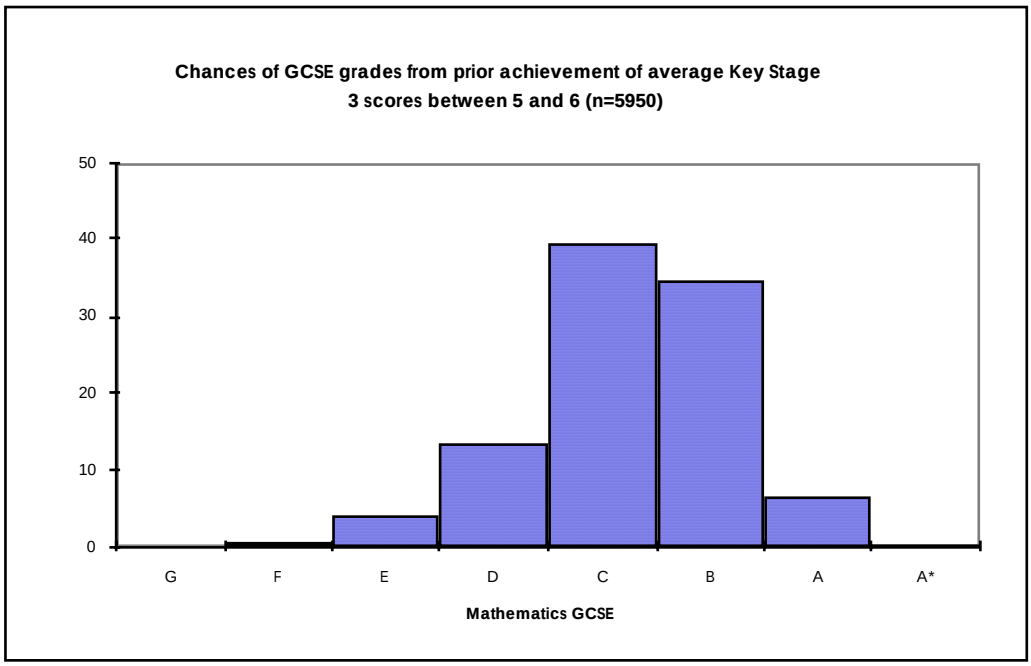
Additionally there are two problems concerning the measurements: unreliability and over-reliance on early measures.

Unreliability of individual predictions

Because individuals are so unpredictable a **point estimate** (suggesting that with a certain starting point a pupil is likely to achieve such and such a

grade on exit) is not wise. Rather than telling pupils a single predicted grade, or having teachers think in terms of a single predicted grade, a better approach is to illustrate the **chances** that a pupil has of achieving all grades from a given starting point (see Figure 4.6).

Figure 4.5 Chances Graph for pupils with an average KS3 score between 5 and 6



To show a pupil the range of grades that were actually obtained from a given starting point is not only more satisfactory statistically but considerably more motivating since, rather than appearing to close doors, it opens up opportunities and possibilities. The fact that early achievement only predicts about half the variation in subsequent achievement means that almost any grade or level is obtainable subsequently from the usual range of starting points.

Over-reliance on early measures

There is a need to maintain the motivation of all pupils and not write off any pupils as unlikely to achieve. The existence of low scores on the prior attainment measure could present problems here. There could be a danger that under achievement at one point in a pupil's academic career could become a self-fulfilling prophecy. Although the concept of under-achievement is somewhat contentious since it implies we can know the true level of achievement against which 'under' achievement is judged, nevertheless most teachers recognise that some pupils fail to work adequately or effectively for some periods of time, perhaps because they are confused or perhaps for other reasons. In this sense there is agreement on 'under achievement'. The use of the Chances graphs counteracts this danger to a certain extent.

4.4 Use of additional tests and analyses

Another approach to under-achievement is to compare scores on curriculum-free '**aptitude'** tests with scores on the tests of curriculum-embedded '**achievement**' tests (i.e. examinations and key stage tests).

However, since there is no guarantee that a pupil is trying to achieve on the curriculum-free test any more than on the taught work, *low scores* are always somewhat in doubt. High scores, however, cannot be obtained without certain skills and are therefore likely to be more reliable and informative.

The danger of low expectations of pupils with low intake scores is best counteracted by the use of the chances graphs. However, altering patterns of effort may be no easier, in reality, than altering patterns of developed aptitudes (Carroll, 1963).

Aptitude tests should not be thought of as measuring innate or immutable aptitudes, they are simply another measure of current, developed aptitudes, filtered through the levels of motivation and application mustered by the pupil at the time of the test. However, they can be useful.

The addition of a curriculum-free test of developed abilities may sometimes

- improve prediction allowing better measures of value-added to be computed;
- allow schools to test their concerns regarding early over-achievement that might be depressing value-added scores;
- help teachers to identify students who might reasonably be expected to be achieving more highly on the curriculum-embedded (achievement) tests;
- give an alternative perspective on a major aim of value-added analyses: fair comparisons between schools;
- be used in the same year as the end of key stage tests in order to overcome the problem of pupils who have only been in the school for part of the key stage (a particular problem the four year span of Key Stage 2).
- provide fair comparisons in a one-year cycle that is useful for schools waiting for value-added measures to become available.

Yet another view of the performance of pupils in various subjects can be obtained by the computation of 'Relative Ratings' comparing pupils' achievement in one subject with their achievements in other subjects. This procedure was developed in Scotland and has been used there in providing indicators for schools since 1991 (Kelly, 1976). Oversimplified versions that do not adjust for differences between subjects are open to question. Since the mathematics behind the calculations is not simple and large datasets are needed to establish differences between subjects, this kind of information is generally provided centrally for schools.

4.5 Need for staff development and INSET

A major factor determining the use made of value-added information will be the presence in the school of someone with considerable confidence in interpreting the data. Such persons are more likely to be found in secondary schools than in primary schools giving cause for serious concern about the use of value-added information in primary schools without

external support to interpret the data. This issue is further considered in Section 6.

As shown in Section 2, there was a strong demand for INSET and for telephone helplines from both primary and secondary schools and, importantly, an experiment showed that INSET can increase the perceived need for help and the perceived usefulness of the value-added information[12].

The development of software, including videos, computer discs and compact discs, with examples of the use being made of value-added by teachers and headteachers, would be helpful to schools that will be entering this area without prior experience.

An important use will be to find methods of teaching and learning that improve achievement. However, there is a danger in simply relating changes in value-added to concurrent changes in practice. A new practice could be introduced and the value-added might become worse. But perhaps it would have become worse still without the new practice. "Correlation is not causation" and unfortunately **a value-added framework will not yield simple and easy evidence regarding effectiveness.** To ensure that seemingly effective practices are in fact the cause of observed results requires additional ways of gathering evidence, ideally by properly controlled experimentation, ie evidence-based education. Certainly careful evaluations are needed of practices before they are strongly recommended, particularly if it would be costly in effort or resources to implement them. Value-added data, year on year, will be helpful, but will still leave much unexplained.

The use of teachers' assessments of pupils

The use of teacher assessment has been fully supported by SCAA. Moreover parents appreciate that teachers know their pupils in greater detail than the information that can be gleaned in a brief period of testing. Parents therefore appreciate teacher assessment as drawing on a broader body of knowledge of a pupil than a single test. It is however quite clear that in some cases teacher assessment is simply geared to the test results so that the correlation is practically perfect between teacher assigned levels and the test results (Second Primary Technical report). When this is the case, the purpose of teacher assessment, its breadth and distinctiveness, is lost. The value and status of assessments made by teachers may be best enhanced by reporting these assessments separately.

In order to make fair comparisons, teacher assessments cannot be used in the value-added calculations for a national system. They cannot be made strictly comparable from school to school since they involve different teachers in every school applying criteria in different ways, and they are subject to possible biases. If teacher assessments were the basis for value-added assessments they would, to a large extent, be under the control of the

---

[12] INSET was randomly assigned with the results indicated. See Tymms' Primary Report 2.

teacher and inevitably there would be temptations towards biasing results to make the school look as good as possible. The value-added system should depend purely on 'blind' [13] externally marked components. The teacher assessment components and coursework should be separately reported.

Teacher assessments can often be better predictors of subsequent achievement than are tests, though the accuracy will vary from teacher to teacher. By having them separately reported, research can be undertaken which might confirm that they are better predictors of subsequent achievement than are tests. Furthermore employers may recognise that, for some kinds of jobs, teacher assessments will provide better prediction of job performance than will test scores. This will almost certainly be the case because teachers' knowledge of the pupils is more thorough than can be gleaned from a test but also because the tasks on which they base that knowledge can be extended in time and be more directly comparable to the kind of tasks that have to be accomplished in a job.

Although teacher assessments cannot be used in a formal value-added system, they can be used as part of the internal-to-the school, value-added analyses. *Before* the outcome data becomes available teacher assessments can be made and either used alone on the horizontal axis or combined with the prior achievement levels to form a composite scale. When the outcome measures become available, the value-added scores can be re-calculated using either scale. The predictions may have improved and the changes in value-added scores for individual pupils will generate discussion both about the pupil's progress and about the basis of the teacher's insights. Teacher assessments could also form the outcomes for monitoring progress throughout the Key Stage. To check if individual teachers are making accurate predictions will be simple once the independently assessed outcomes become available. It may be the case that, because they know their pupils better, teachers who can make accurate predictions can also obtain greater rates of progress. This is a topic in need of research, along with the impact of tracking and monitoring processes.

In concluding this section on internal school use, it can be suggested that the use of value-added data within schools has the potential to stimulate discussions and investigations but the data will need to be interpreted with care and caution. This kind of interpretation will be aided by the indication of margins of error (although these cannot be interpreted rigidly), the use of clear and simple presentations, and, particularly, by the provision of

---

[13] Marking is 'blind' if it is undertaken by someone with no knowledge of the pupil's name or school, (information that conveys gender, ethnicity, religion and social class) and no knowledge of any attainments other than the papers being marked. The marks already awarded to a paper should not be known if the paper is re-marked and prior achievement should not be known to the marker. Blind marking is simple quality assurance to ensure that the pupil is judged purely on the test. It involves standard safeguards and is required by a consideration of equal opportunities.

pupil-by-pupil details to every teacher and the collection of two or three years' data before tentative conclusions are framed.

4.7 Value-added data in School Performance Tables

Value-added data could be included in School Performance Tables and/or could be made available to parents, governors and others through local schools, LEAs and equivalent agencies.

The School Performance Tables currently represent only school outputs, with no account taken of the prior attainment that accounts for fifty percent of the variation in the outputs. It was reported in section 2 that almost two thirds of secondary headteachers in the sample wanted value-added data published alongside GCSE results, if there is to be publication.

A 1995 publication "**Value Added In Education"** from the Department for Education stated that

> The secretary of state for education is... committed to supporting work to develop national measures of value-added and, if clear and reliable measures can be established, to publishing them in school and college performance tables."(*DfE 1995. Page 1)*

Since clear and reliable measures can be established, what should be expected from value-added data if such data is to be published or made widely available?

4.8 The information provided by value-added scores for schools

As value-added data becomes available, it will be important to observe the following characteristics in order to avoid over-reliance on any single sets of data:

- value-added measures within schools will vary from subject to subject making it difficult to justify the use of a single indicator of overall achievement;
- value-added measures of all kinds will vary from year to year;
- the sizes of the differences between schools will be seen to be considerably smaller than indicated by raw data. The figures found in the pilots are summarised in Table 4.2 below;
- Because differences between schools on value-added scores are relatively small there will be considerable difficulties in trying to compare two or more schools in a neighbourhood. The differences between them will frequently lie within a large margin of uncertainty. Parents would be ill-advised to rely on a single value-added indicator in choosing between schools.
- There will be some changes in rank orders when based on value-added rather than raw scores, but the use of ranking is still unwise. Rank orders based on value added scores will not remove the fundamental problem that rank orders are inherently undesirable because they fail to convey the *size* of the differences. Any discussion of a school's 'position' in rank-ordered tables is unfortunate. In the centre of the distribution a small change on the indicator can move a school many places. At the outer edges of a distribution even a large difference may have little or no effect on the school's position in a rank order. There is a need to present data so that the *sizes* of differences are the focus of attention, not positions in a list.

Table 4.2. The effect of value-added on the sizes of differences between schools

Phase	Outcome measure	The range of scores within which lie about 68 percent of schools (ie. One standard deviation on either side of the average score)	
		Raw scores	Value-added scores
Primary	Key Stage 2	0.73 of a level	0.46 of a level
Secondary	GCSE-average	1.70 of a grade	0.52 of a grade
Secondary	GCSE-total	20 GCSE points	7.9 GCSE points

Data from the value added national project 1996 pilot studies

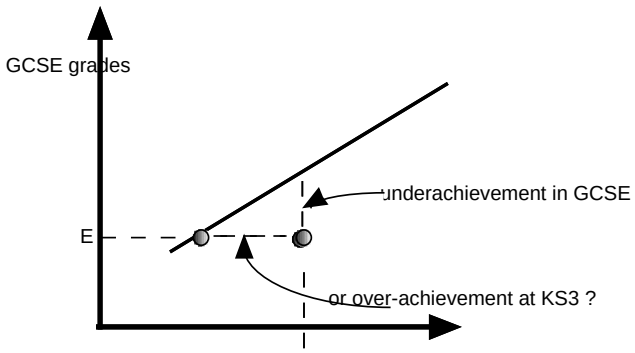
Margins of uncertainty

Much of the variation seen in value-added measures will not be explicable in terms of any actions that the school has or has not taken. If nothing changed from year to year other than the sample of pupils, there would still be variations in the value-added scores. One statistical approach to this inherent variation is to report 'confidence intervals' indicating the extent of this only-to-be-expected variation. Statistical tests and 'error' terms refer to this 'sampling variation'. In Annex D, Tymms shows how simple calculations generally yield 'errors' in line with those from multi-level modelling. They can be used to indicate a margin of uncertainty around results from different departments or around results across several years. However the actual variation may well be larger than the statistically calculated sampling variation. Some of the variation may be attributable to the effects of actions taken by staff and pupils or the effects of school resources or policies, but it will be difficult to pin this down with any certainty, particularly if only one year's results are available. Moreover some of the uncertainty will be due to problems beyond statistical modelling, such as varying motivation of pupils, as indicated in Figure 4.6. There may be pupils for whom the input measures were clearly inappropriate due to lack of effort, illness or stress. In such a case, the value-added may be more of a reflection of earlier inappropriate assessments than current performance levels.

4.9 Choices for school indicators

If value-added indicators are to be included in School Performance Tables, the most important question that arises is *which* value-added indicators: that is, if a whole school indicator is provided, should this be based on an average or total score, and should these aggregations be made across all subjects or only those required by the national curriculum? And should the propensity of some subjects to yield lower grades ('subject difficulties') be taken into account in the method of computation?

Figure 4.6 Unreliability on the input measure makes interpretation difficult



The advantages and disadvantages of average or total GCSE scores were discussed in section 3 and amount to a choice between quality of output regardless of amount (average-GCSE) or amount confounded with quality (total-GCSE). Since one resolution of the issues would be to include both,

a possible format for a whole School Performance Table is provided as Figure 4.7.

Figure 4.7 A possible GCSE School Performance Table containing raw results and a predicted range of results

	GCSE: Whole School Performance Tables						
	2	3	4	**5**	6	**7**	8
:hool	Number of pupils on roll	Percent of roll counted for this table*	Number of GCSE passes per pupil	**Predicted range of value-added indicators: average GCSE points**	Average GCSE points	**Predicted range of value-added indicators: total GCSE points**	Total GCSE points
:hool a							
:hool b							
c.							

Notes:

- Column 2 shows the number of pupils on roll. This is of interest to parents who may prefer large or small schools.
- Columns 3, the percentage of pupils for whom examination results are reported, provides information about the extent to which a school is entering pupils for external examinations. This indicator can discourage a distortion of the indicators by highly selective entry policies[14].
- The figures in the shaded columns (5 and 7) represent the range of variation statistically expected from year to year. If the figure in the next column lies within the predicted range the indicator is in line with the general rate of progress in the national dataset.. The predictions used in this table take into account both the earlier attainments of pupils at key stage 3 and the subjects taken in the GCSE examinations.
- * pupils were counted if they were on the school roll for 70 percent of school year by 1 April 1998
- ** '**predicted' grades** were arrived at for each pupil on the basis of the pupil's prior attainment at key stage 3 and the general pattern of achievement in the GCSE results for similar pupils in other schools. The use of this prediction allows some account to be taken of the differing prior achievement levels in difference schools.

The use of a predicted range enables the concept of only-to-be-expected variation to be conveyed. However, it suffers from the arbitrary decision as to how wide the margin of acceptability should be made.
An alternative approach is represented in Figure 4.8 in which data are presented by curriculum area and the value-added score is directly reported, thus forcing a consideration of the magnitude represented, rather than relying on arbitrary significance levels. In Figure 4.9,

[14]It is generally recognised that publication of indicators about a system produces responses from the system that may sometimes be designed more to alter the indicator than to alter the reality that the indicator is supposed to represent. One way to avoid distortions is to report several indicators that check on the possible sources of distortion.

mathematics/science entries are reported separately from humanities entries and separately from foreign language entries, performing arts, vocational areas, design and technology, etc.
The difference in performance between boys and girls has been a topic of concern, perhaps exaggerated but nevertheless real. By including columns in which boys are compared only with boys and girls' value-added compared only with girls', the extent to which there are differences, both within a school and generally, will be apparent. Furthermore if one of the columns is empty it signals a single sex school which is again a matter of information of interest to parents. This kind of detailed School Performance Table differentiated by curriculum group and by gender might be provided on a local basis rather than nationally.

Figure 4.8 A School Performance Table With (A) Differentiation By Curriculum Group And (B) Gender Differences Taken Into Account

1	2	3	4	5	6	7	8	9	10
School	Math - Science (MS) entries as a percent of all GCSE entries	Av. Grade in MS subjects	Difference between predicted and average MS grade: boys only	Difference between predicted and average MS grade: girls only	humanities entries as a percent of all GCSE entries GCSE	Av. Grade in humanities	Difference between predicted and average humanities grade: boys only	Difference between predicted and average humanities grade: girls only	etc.

An important point with respect to the provision of value-added data in Table 4.10 is the way in which value-added is described in Columns 4 and 5 as a difference score, the difference between the predicted average grade and the actual average grade. This 'difference' concept is probably easier to convey than the concept of positive and negative value-added from regression line diagrams. Those not keen to follow the methodology can simply accept that the prediction is based on prior attainment and therefore has taken prior attainment levels into account.

However the most important feature of this method of presenting the value-added is that it focuses attention on the size of the discrepancy. Differences which are a mere fraction of a grade might, with very large numbers, be statistically significant but not of substantial importance. The difference indicator is straightforward and readily interpretable, though sample sizes must be considered.

As long as the differences between predicted and average have been based on **prediction within each subject or syllabus** there will be no penalty for schools entering pupils for 'difficult' subjects or syllabuses.
Parents may be concerned to choose schools with certain strengths, strengths that relate to the interests and capabilities of their children. For these reasons parents may wish to see the value-added differentiated according to different areas of the curriculum as in Figure 4.8.
If value-added for individual subjects were to be provided this would frequently be tantamount to undertaking personnel work in public, in subjects taught by single teachers. This would probably not be acceptable.

Age Cohorts or Year cohorts

The Performance Tables are based on age groups whereas value-added will be based on year groups, as is currently the case with key stages 1 to 3. This will need resolution.

In conclusion

It is likely that the publication of carefully collected information is here to stay. The solution is to ensure, by the choice of indicators, that the impact is as beneficial as possible. This concern is addressed in Section 6. Prior to that, the way in which a value-added system could economically produce acceptable indicators is the issue addressed in Section 5. Only if the provision of *good* data is organisationally feasible is there a basis for a national system.

SECTION 5: ORGANISATIONAL MODELS

The central question for this section is: what would need to happen in order to have a national system in place very shortly? Immediate strategies are needed, following which any system will undergo further development and elaboration, particularly if assisted by careful evaluation. The approach taken is to build on existing structures for the provision of timely information in the Stage 1 feedback to schools for internal use, and to consider the addition to national data collection to provide the datasets for Stage 2 analyses, including exploratory data analysis.

The possibility of value-added data becoming publicly available, whether or not it is published nationally, makes the examination and assessment systems even more 'high stakes' than was formerly the case for schools. Additional safeguards and quality assurance procedures must now be built into the system. This issue is therefore addressed in this section, particularly since the procedures adopted for marking will influence the feasibility of implementing adequate quality assurance systems.

The existence of unique pupil identifiers and open Management Information Systems, in which files are fully accessible and under the control of the user, would make all this work considerably more efficient: less costly and less error-prone. The practical advantages of unique identifiers have been recognised in an internal DfEE study (the Flint review), which recommended their adoption for the generation and maintenance of standard pupil records to serve a wide range of administrative and statistical purposes.

There may be concerns about the use of a unique pupil identification number, a facility that would greatly reduce the problems of running a national system. Such a number would simply be an addition to the unique identification numbers already in use for health, national insurance and employment. Furthermore, there is a compulsory schooling requirement and yet no way adequately to monitor its implementation. A proportion of teenagers may have simply dropped out of secondary schools without any consequences. A unique identification could be a positive benefit in more ways than simply speeding up value-added measures.

Terminology: The 'external marking agencies' (EMAs) manage the marking processes for the assessments at the end of key stages two and three (KS2 and KS3). A similar role is undertaken by the examination boards at the end of compulsory schooling (the GCSE examinations). The term 'grade' is used to indicate the awarded examination *grade* for GCSE - the equivalent term for the end key stage tests is *level*. Reference will be made to national data collection agencies. These could be any organisation with the processing capacity and quality assurance procedures to ensure the integrity of the data received. Some of the procedures proposed below envisage that the agencies will collect the data for the compilation of national aggregates, for dissemination to LEAs and schools, and possibly for national publication. A national data collection agency for key stages

two and three is already in existence. The data collection functions at GCSE are undertaken by the examinations boards and by Bath University. Whether the data collection agency or agencies should archive the data for long-term retention is not discussed. The Economic and Social Research Council has a Data Archive facility with long experience of maintaining files and making data available to researchers.
Value-added procedures for GCSE examinations are considered first, followed by procedures for the situation in which the output measure is an end of key stage test.

5.1 Data capture for a GCSE value-added system

A system which could be put into practice immediately is shown in Figures 5.1 and described below. It could only be used, however, by schools that had retained the average KS3 score for each pupil. The average KS3 score can form the basis for value-added calculations for each GCSE syllabus, as illustrated in the secondary technical reports.
A key to this system working efficiently is that when pupils are registered for GCSE examinations, their prior attainment KS3 score is included in the registration (see pupil registration in Figure 5.1[15]). Following pupil registration, schools would receive either an examination number for each candidate or a set of candidate entry records for each pupil incorporating the unique pupil identifier if one had been assigned.
After the administration of the GCSE examinations (Time 2 in Figure 5.1), the usual procedures for translating marks into grades are followed by the examinations boards. Once the results have been awarded, *and not before*, the prior attainment data can be linked with the grades. Highly unexpected results should be subject to *blind* re-marking, with the marker unaware of the direction of the discrepancy from the results predicted by prior attainment. **It is important that pupils are judged only by their current performance and their assessment is not affected by their prior attainment.** Pupils can change their performance dramatically and can and do obtain almost every grade from almost every starting point. Prior achievement only accounts for about 50 percent of the variation in outcomes and it is important that the system of assessment *each time* enables every pupil to be marked *without any bias introduced by a knowledge of previous attainment levels.*
Once grades are agreed, the prior attainment and the grades are matched for a particular syllabus, so that the value-added regression line can be computed for that syllabus.

5.2 Data capture for an A-level value-added system

---

[15] A procedure suggested to the Value Added Advisory Group by John Gardner, former Chief Statistician with the DfEE.

The procedures just described for GCSE examinations can be applied to A-levels immediately. The predicted grades will be computed from average-GCSE scores submitted by schools and colleges at the time of registration for each syllabus. The well-established differences in regression lines for different A-level subjects and syllabuses (Fitz-Gibbon and Vincent, 1994) do not pose a problem if like is compared with like: value-added is computed per syllabus, thus matching the analysis to the process that produces the data.

5.3 Data capture for value-added systems for end of Key Stage tests

For the calculation of value-added at the end of KS2 and KS3, schools will need to have acquired and retained, at the very least, the average attainment from the end of the previous key stage. Thus in secondary schools the attainment at the end of KS2 would be used as the basis for value-added calculations for KS3, although it was not part of this project to investigate this possibility. In primary schools, when the KS1 data is retained, it will be possible to have the value-added scores computed to either an intermediate test halfway through KS2 or to the end of Key Stage 2 tests. It may be necessary for the data collection agency to collect the data nationally at the point of registration of pupils for the key stage tests, a procedure directly analogous to that proposed above for GCSE. Once collected, the prior attainment data would be used to create value-added regression lines based on all the available data, as illustrated in Figure 5.2, or if national collection were not followed, a sampling approach to collection could be sufficient to provide 'national' regression lines.

5.4 Stage 1 feedback to schools, for internal use.

Having matched, for every pupil, the current year's data with prior attainment, syllabus by syllabus for examinations and subject by subject for the end of key stage tests, there could be two main ways of reporting the results, the regression-line option or the ready-made value-added tables option.

Either:

The regression-line option. The regression lines could be made available to all schools, by mail, by national publication or on disks. Those without sufficient skills to proceed with the calculations using a standard spreadsheet could be provided with appropriate software or could obtain external help from their LEA or other groups.

Or:

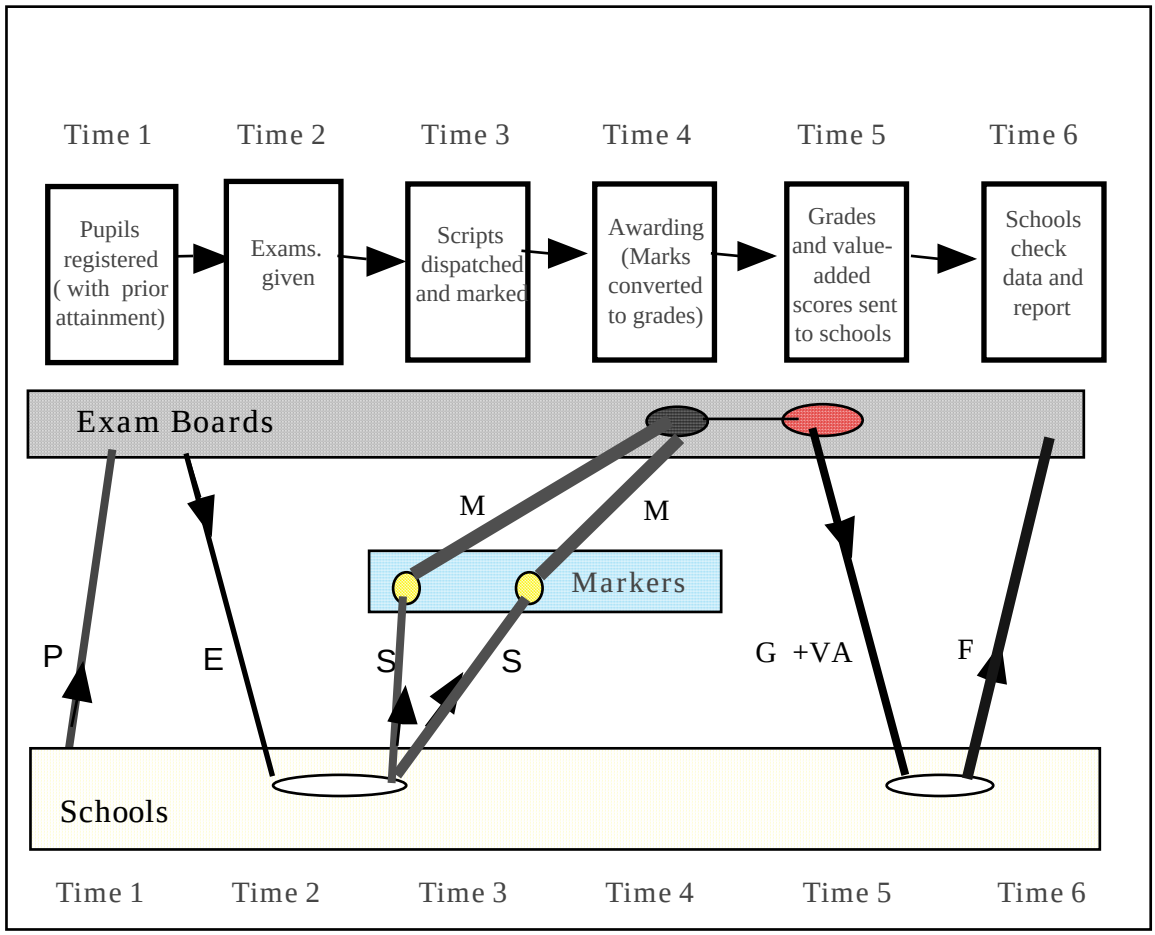
The value-added tables option.

Value-added scores, for each pupil, could be provided in tables alongside the test or examination results (cf. Table 4.1).

The provision of regression lines is more economical but puts greater onus on the schools or other groups. Initially there would be some schools without access to the necessary skills. Thus, in the short term, the provision of 'official' value-added tables might be the method of choice. The systems illustrated in Figures 5.1, for GCSE and A-levels, and 5.2 for end of key stage tests, could be implemented immediately and would provide value-added information *for those schools that have retained prior attainment data.*

Primary schools that have baseline data for pupils on entry, will be able to have value-added scores when key stage 1 data become available, but these will only be valuable if the baseline scores correlate well with KS1 attainments.

Figure 5.1 Building value-added into existing arrangements : GCSEs and A-levels.



KEY: Information exchanged

P — Pupil registration (including prior attainment)

E — Exam. papers

S — Scripts

M — Marks

G +VA — Grades PLUS value-added scores provided directly to schools

F — Final list of results, grades and value-added scores

KEY: Processes

Exam. administration

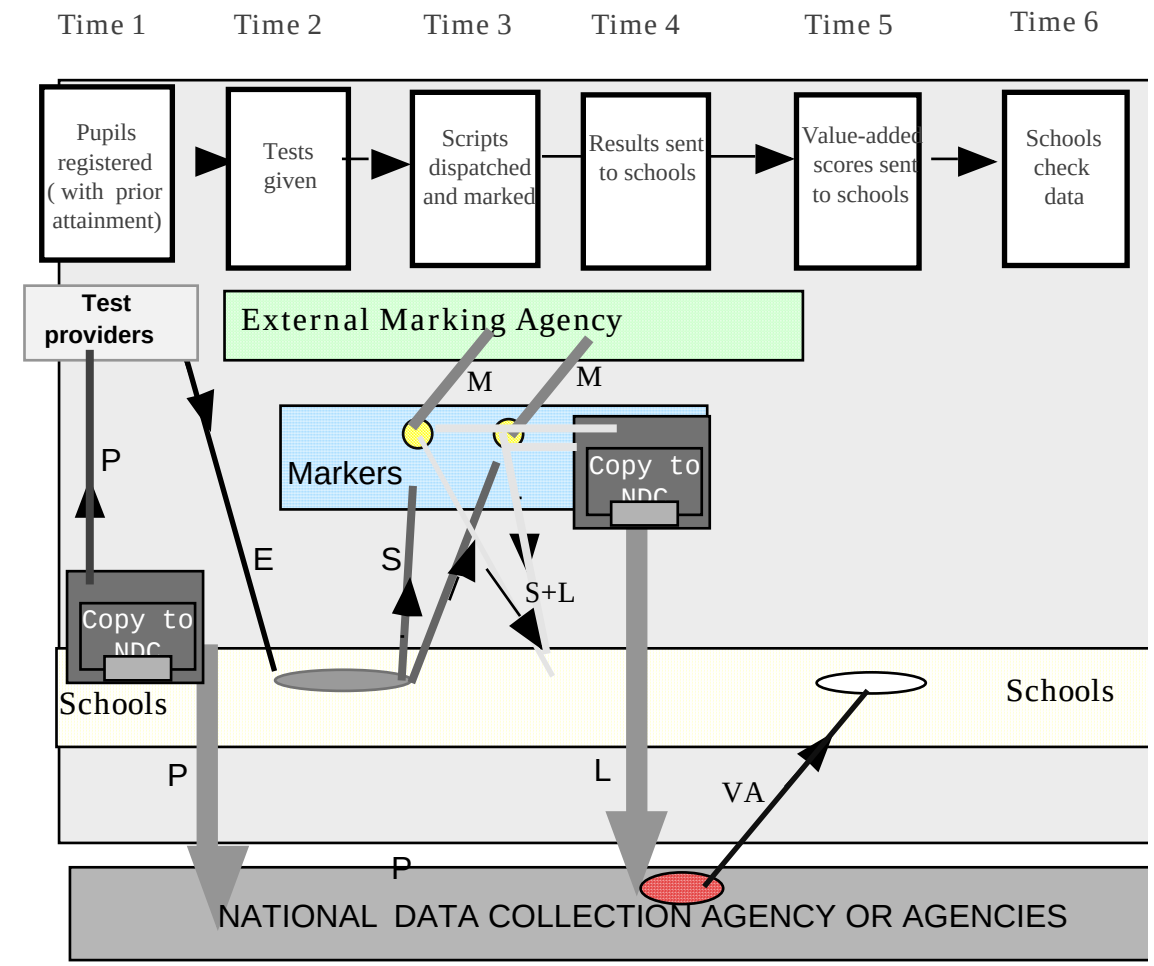
Marking

Awarding process

Grades linked with prior attainment

Final checking of all data

Figure 5.2 KEY STAGE TESTING
Providing value-added data to schools from key stage testing



KEY: Information exchanged

P — Pupil registration (including prior attainment)
E — Test. papers and EDI lists of prior attainment
S — Scripts and EDI lists
L — Levels
VA — Value-added tables provided directly to schools

KEY: Processes

Test. administration
Marking
Levels linked with prior attainment
Final checking of all data

5.5 Stage 2 reporting, possibly for publication.

If publishable national data are desired, then additional procedures are recommended, as described in Section 3. For example, further checking of the data will be necessary, and more elaborate analyses can be carried out to ensure that atypical schools are evaluated as fairly as possible. Even if there no national publication schools will be required to share performance data with inspectors and governors, so that the difference between publication and public availability is not as great as might appear at first sight.

A minimal second stage

A minimal second stage, using only the data already captured, can be accomplished for GCSE by copying three of the intermediate datasets to a data collection agency, as illustrated in Figure 5.3 (GCSEs and A-levels). For end of key stage tests the national data collection agency already receives the three equivalent datasets, as shown in Figure 5.2. Given the desirable use of electronic data interchange procedures and unique pupil identifiers, national data collection at GCSE would require no more than the copying of datasets. Furthermore, this simple and economical approach can also provide the data for monitoring the assessment process, serving as a basis for quality assurance procedures built into the examining/testing programs.

For the external GCSE examinations, the three points at which datasets need to be copied are:

- **At pupil registration**, schools send pupil lists not only to the examination board but also to a data collection agency;
- **When the marks are sent** by markers to the examination boards they can also be copied to a data collection agency;
- **When the grades and value-added scores are sent to schools** they can be copied also to a data collection agency.

The proposed data collection agency for GCSE would then be in possession of the prior attainment measures, the marks awarded and the translation of these marks into grades. At the request of the assessment authority, checks can be made where marks and levels/grades do not relate simply, or a random subset of the data can be audited. These checks should be additional to blind re-marking of a randomly selected set of scripts in order to establish inter-marker reliability and conduct generalizability studies of the kind advocated by Johnson (1996). These kinds of quality assurance procedures become important as the consequences that might follow from value-added analyses become salient and of concern to parents, pupils, teachers, governors and all concerned with the reputation and welfare of schools, individually or collectively.

There are a number of options within the design presented above. If the examination boards cannot provide a value-added analysis for each GCSE syllabus at a reasonable cost, this could be done centrally through the

proposed data collection agency, public or private, and commissioned each year if necessary. However, given the existing channels of communication between the examination boards and schools, and the fairly trivial nature of computing a regression line and reporting the value-added scores, it would seem most sensible to use these existing channels and relationships. The value-added calculations for KS2 and KS3 would be carried out by the existing data collection agency. In all cases the additional costs should be proportionate to the small increase in effort. In fact, since prior attainment measures provide statistically 'predicted' grades for GCSE once the current year's data has been analysed, they could replace the use of teachers' predicted grades. The value-added process could thus represent almost no additional cost for GCSE.

Additional safeguards required for central data collection

Certain safeguards would need to be considered, such as the following:

- the raw data used to calculate value-added should be double checked by schools before being sent for national recording and publication;
- the Data Protection Act would apply and all parties would need to be registered and be prepared to comply with the requirements;
- pupils' unique educational identifiers should be used for national storage of the data, thus anonymising the datasets;
- files in which unique educational identifiers are recorded, linked to real names, should be kept secure and separate from data files.

In a minimal second stage, as suggested in Section 3.2, a variety of analyses would be applied to the available data, starting with that thorough exploration of the pattern of each variable generally referred to as 'exploratory data analysis' (Tukey, 1977; Hartwig and Dearing, 1979) or 'initial data analysis' (Chatfield, 1985). Smoothing procedures (cf figure 6.1), and restricting regression plots to segments representing the actual range in the data (cf figure 3.2), illustrate two ways in which further insights can be gained, in particular with regard to the sizes and educational significance of differences.

An extended second stage

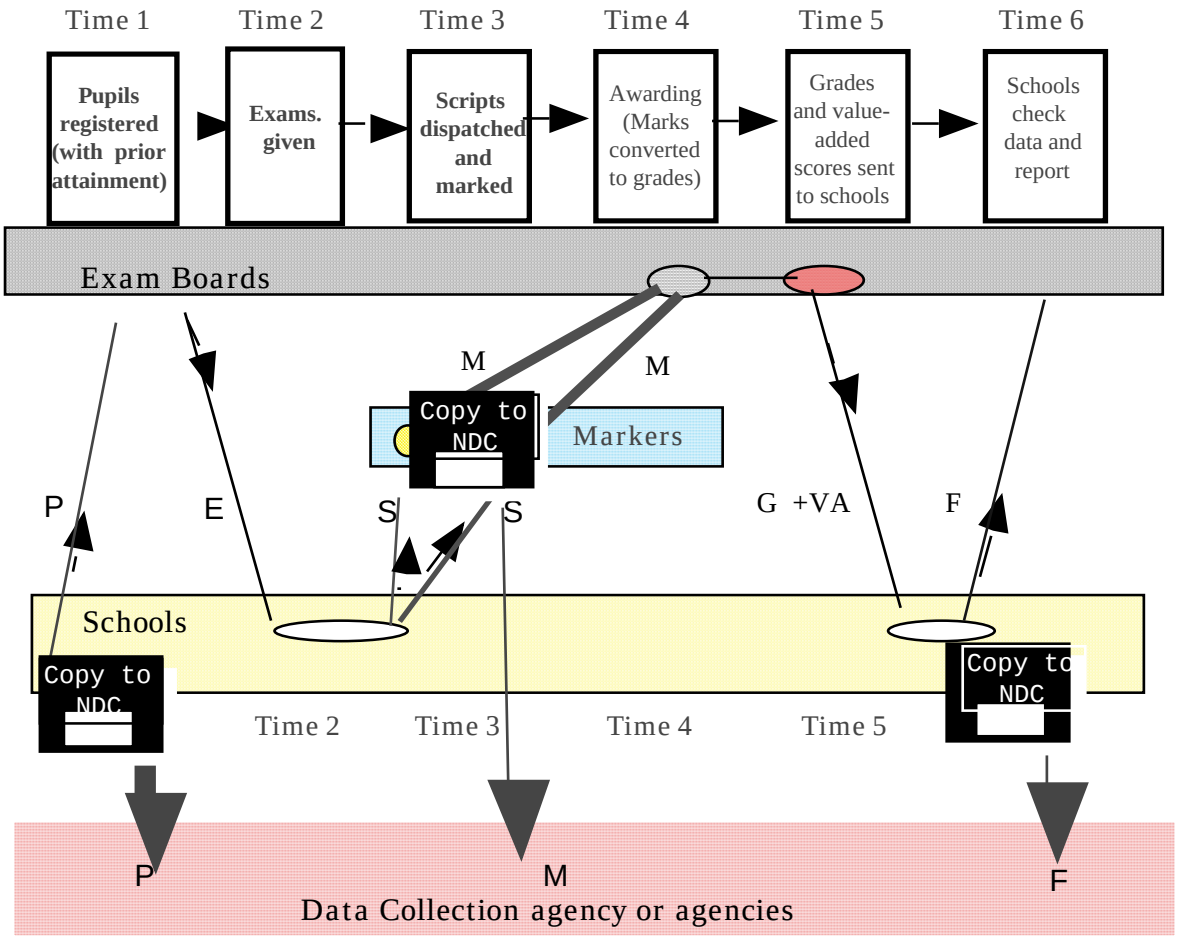
An extended second stage would include additional variables such as English as a second language, sex and social measures, as discussed in Section 3.3. Additional data that might suggest hypotheses about performance could include expenditure per pupil, and pupil-teacher ratios. Methods of teaching and learning and alterable school policies would probably provide more useful information for school improvement.
By having a regional or local authority dimension for further exploratory analysis and support work in schools, the insights provided by a thorough knowledge of local conditions could be brought to bear. For example the inclusion of employment rates in the datasets, the opening of a new housing complex, or relating bus routes to attendance patterns, could all be influential in interpreting the performance of schools. This level of contact

with realities behind the data would scarcely be practical on a national basis.

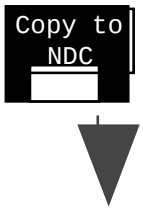
5.6 The impact of publication.

If value-added data is to be prepared for addition to the school performance tables, or to be available publicly in some other form, then additional efforts need to be made to have entirely standardised means of dealing with factors that affect the outcomes, to allow any comparisons to be as fair as possible to every school. This requirement will be particularly difficult to make with respect to attendance data.

Figure 5.3 PUBLIC EXAMINATIONS:
Simply adding three copying procedures to existing arrangements can provide
- **monitoring of the examining process**
- **a national database of value-added information**



KEY: as in Figure 5.1 with the addition of:



= indicates that an exact copy is provided tothe NDC, on disk

NOTES:

* Electronic Data Interchange is deemed highly important if a system is to run promptly and efficiently

* the Data Collection agency or agencies will receive prior attainment, marks of random scripts, grades and value-added for each syllabus/assessment process. This enables monitoring of the inter-marker reliability and the consistency of standards against prior attainment levels

If the value-added indicators are for local use, then the difficult issue of pupil attendance could be less critical. If, for example, the value-added scores related only to a proportion of pupils with some defined level of

'adequate attendance', parents could interpret the data as the likely effect of the school on those pupils who had adequate attendance.[16] However, for public accountability the issue of the exclusion from the calculation of non-reported pupils would be a matter of concern.

In short, for national publication, in contradistinction to internal school use, the issue of pupil attendance becomes extremely important, along with the checking of data by schools.

5.7 Non-reported pupils, attendance and mobility

Schools cannot teach pupils who are absent. Nor can they be responsible for the progress of a pupil over a key stage who has arrived in the school only a month or two before the end of key stage tests or examinations. Furthermore, excluding pupils, or removing them from lists in view of poor attendance records, are ways in which the value-added indicators could be influenced. There must thus be an agreed system for dealing with the question of which pupils should be counted in a value-added indicator. The danger is that the pressures of publication will motivate schools towards providing the most favourable view of the data possible. If schools are to be held accountable for pupils on their roll, there must be uniform and objective methods for collecting the relevant data. A review of current methods for reporting attendance data may therefore be required prior to publication of value-added indicators.

Attendance, like time-on-task, may turn out not to have the powerful impact that would seem likely. In the data from primary schools in Avon, for example, mobility did not relate strongly to the progress made by individual pupils but in the secondary pilot study the pupils who were recorded for KS3 but not for GCSE results, had KS3 averages almost a level lower than those who were retained in the school and entered for GCSEs.

The information stored for value-added purposes would ideally be the percentage of lessons for which the pupil was present but, depending upon the extent of computerisation of attendance records, obtaining such data could represent a high effort, high cost approach.

A simpler solution than storing records of attendance and pupil turnover would be to ensure that every key stage test result was linked with a school identification and that this school identification was included in the pupil records which moved with the pupils when they changed schools. Three indices could then be defined : attrition, in-flow and mobility. Pupils present at intake but not at the outcome represent 'attrition', the loss of pupils over time. Pupils who have prior achievement, intake data from a different school represent 'in-flow' to the school. In-flow added to attrition can provide an indicator of mobility among pupils. There would be

---

[16] In the Tennessee Value Added project (probably the only system comparable in scale to the one under discussion here) 'adequate attendance is defined as being present for 80 percent of the year's total school days. (Sanders and Horn, 1995).

problems with various patterns of school transfer, such as from middle to secondary schools at the ages of 12 and 13, but adaptation would not be difficult. The virtue would be that the mobility data would be linked with the data needed for the value-added analyses automatically, and it could be incorporated into the registration process. Consequently, the attrition, in-flow and mobility indices would not require separate data collection. A factor that these indices would not pick up would be the number of changes of school a pupil had made during the key stage. Thus the indices would represent proxies for mobility. Such data might nevertheless have some explanatory power particularly at the level of the school and should certainly feature in the Stage 2 data analyses.

Another solution would be to have an 'allowance' of 10 percent of pupils whom the school could choose not to count for publication. A blanket 10 percent would mean that a school could accept up to 10 percent of its enrolment from 'difficult' pupils (e.g. previously excluded pupils) without fear that this would automatically damage the school's position in academic performance tables. This would encourage all schools to accept some difficult pupils, an important factor in preventing 'sink schools'. However, the choice of which pupils to remove from any public reports should be announced at the time that examination entries are prepared - - - not after the receipt of results. Outcomes for the non-reported 10 percent could be evaluated by other indicators and by inspection. This, and the funding associated with the pupils, should ensure attention to these pupils and free the creative approaches teachers can bring to bear on the varied and specific kinds of problems that some pupils present. There will always be pupils for whom, at times, a press for achievement is counter-productive, and the system must allow for this within reason. Mental and physical illness, accidents, death in the family and other stresses need to be accommodated by any system.

In the use of the 10 per cent discretionary non-reporting, the safeguards should be that

(i) the reasons for non-reporting must be explained to inspectors;
(ii) the pupils whose data is not to be reported must be identified at the time of pupil registration, not after results become available.

Yet another solution would be to have schools choose the percentage of pupils whose results should not be counted due to special reasons such as long absence, stress or other factors, but then have that percentage included in the performance tables and/or be subject to audits.

The method chosen for dealing with the problems of truancy and mobility will probably be a source of concern and discussion for several years. If for each pupil the percentage attendance is recorded various methods of analysis can be considered. One approach would be to follow the example of Tennessee and not count any pupil with less than about an 80 per cent attendance record. In general, however, the drawing of arbitrary lines through a continuous distribution leads to various attempts to alter the data

near the boundary. The important aim is for accurate data so that schools and LEAs or equivalent organisations can learn how best to manage problems of truancy and how best to provide for pupils who, perhaps of necessity, move from school to school.

Special Educational Needs

The issue of including statemented pupils in the value-added system is a complex one and must probably await the development of a research-based and complex system to reflect all matters of concern. One issue that such research would have to address is that the available scales *f* for both input and output measures *f* may be too poorly differentiated to yield useful data for many pupils.

5.8 Value-added for school improvement

The purpose behind the development of indicator systems is to improve the system. It is therefore highly germane to consider how value-added data might be used for school improvement. Some indications were given in Section 4, concentrating mainly on the use of value-added by schools. Considered here is the way in which the use of the data nationally can be designed to lead to school improvement. It can certainly be argued that simply the provision of value-added feedback can be expected to lead to improvement (eg Fitz-Gibbon, 1996) but further efforts may also help.

The best way to use the potential of the large national datasets would probably be by letting a number of contracts for exploratory data analysis on the understanding that the effectiveness of the expenditure would be evaluated by the extent to which the information provided to schools was useful and was effective in improving their performance. A number of organisations bidding for such contracts might form networks with schools to provide a flow of information and support to improve achievement. (It is likely that these networks would quickly discover that the important factors in improving achievement are proximal variables such as school policies and classroom practices, particularly the latter if current research syntheses are correct, eg Wang, Haertel, and Walburg, 1993).

Provision of the data to not just one agency but several might be desirable in that this would put each agency into competition with the others in order to provide the most insightful analysis, the most defensible and the most informative, at the most competitive price.

The evaluation of these analyses should use several empirical criteria. This empirical validation of the models developed is the touchstone; arguments and mathematics will not resolve the issues. One would be the extent to which the data analysts were effective in predicting the subsequent year's results.

Another highly important criterion would be the extent to which the insights could actually be used to improve the system. Improvement may be difficult, but as a minimum it would be desirable to do no harm. The feedback of misleading information is likely to be damaging to schools, costly in wasted efforts.

A 'customer-in-control' model.

Instead of contracts let by a central authority, another approach would be simply to make the datasets available (from the data collection agency or agencies) on a very wide basis and leave it to schools, LEAs and others to commission research on aspects of the data about which they had concerns. In this way the exploratory data analysis need not be funded centrally. The funds instead could be put into schools themselves to enable them to seek such support and help as they may wish in order to solve problems and improve their effectiveness. Just keeping the money for the school might even prove the most effective strategy. Since many schools are already receiving value-added information they would be discriminating customers.

Although the term 'improve' has been used liberally in the foregoing discussion,. this is not meant to imply that schools are in any way currently deficient. By the provision of data on value-added, schools will be able to see if other institutions are getting better results with similar pupils than they are themselves. This will lead to a certain amount of bench marking by teachers, departments, schools and LEAs and this alone should lead to improvements in the system. The system as a whole can improve if the less effective departments move closer to the range of the most effective departments. This process would perhaps be encouraged by publication of the top 10 percent of value-added results in each syllabus (for GCSE) or subject (for KS2 or KS3).

5.9 Uses to be made of regression lines

As soon as prior attainment data are matched with levels or grades and regression lines produced, we have a method of comparing the apparent, statistically implied, 'difficulty' of subjects or syllabuses and external marking agencies. Thus for GCSE the assessment authority can immediately report any adjustments that need to be made in order to equate, post hoc, the difficulties of different syllabuses within the same subject. Information can be published which will be useful for admissions officers and employers comparing the differences in difficulty of different subjects. This same information will save some teachers from being misjudged because their pupils were entered for 'difficult' subjects or syllabuses.

At GCSE the appropriate way to compare like with like is to create regression lines for syllabuses not subjects, since marking and grading represent a process for each particular syllabus. This will enable immediate checks to be made on the equivalence of syllabuses, a matter of considerable concern in schools. Any neglect of differential syllabus or subject difficulties, or of examination board differences, would cause concern to a greater extent as the 'stakes' became higher, particularly since it seems that each year as few as 10 per cent of syllabuses are currently studied for comparability from board to board (Dearing, 1996). It is

important to individuals, schools and nationally that syllabuses and subjects can be chosen for reasons of content and suitability for purpose rather than with an eye to the likely value-added indicators.

5.10 Roles of such organisations as Local Education Authorities

Local education authorities (LEAs) were assigned a monitoring role in Circular 9/88 issued by the Department of Education and Science. In order to monitor schools intelligently, they need to have access to the value-added data. They are very likely to have a role in interpreting value-added data on schools' behalf, particularly with primary schools. The same would apply to agencies with responsibilities for other groups of schools, such as the Funding Agency for Schools. However these organisations and LEAs should be considered in part responsible for the impact that schools can have on the attainment of pupils. It is their role to support schools and enhance their effectiveness. These agencies should not, therefore, be involved in data collection and the primary analysis, but they will need access to the data.

If there is national data collection, it will be perfectly feasible to create value-added indicators for LEAs and similar groups. Indeed all organisations and individuals claiming to know how schools can be improved will have the chance to show that their advice is effective. It must be recognised, however, that due to the variable nature of value-added from year to year the data will be difficult to interpret. The best evidence for effectiveness will arise from properly controlled interventions among randomly equivalent groups of schools (ie groups of schools randomly assigned to different interventions as is the basic underlying principle of what is coming to be called 'evidence-based' policy).

5.11 Conclusions

An impressive monitoring-with-feedback, value-added system is feasible and can be economically designed by building minor additional processes into existing testing and examination procedures. In addition to providing schools with regular feedback on the relative progress of each pupil, the additional processes can improve the quality assurance provided in the examination process. Such improvements are now essential in view of the increasing consequences associated with value-added monitoring.

These developments highlight the need for an infrastructure that enables

- data to be exchanged electronically
- the procedure of matching datasets to be conducted on the basis of unique pupil identification numbers.

Without such an infrastructure, the costs will be considerable, the errors more numerous and the timeliness of all the work may be inadequate.

If value-added data are to be used for publication, additional safeguards will be required such as secure systems of checking test/examination results and recording attendance/mobility or any data that can result in non-reported value-added measures.

SECTION 6: WIDER ISSUES

The benefits of having a value-added system are widely appreciated. It will be wise to consider possibly harmful effects and, towards the end of this section, to consider the essential role of government agencies in the setting-up of such systems.

In a 1995 article entitled 'On the unintended consequences of publishing performance data in the public sector' Peter Smith, Professor of Economics at the University of York, wrote

> '*Most performance indicator systems will ... fail unless serious consideration is given to the deficiencies described in this paper'.*

He also noted that

> '*the ultimate criterion for judging the usefulness of a (Performance Indicator). scheme is the magnitude of its benefits in relation to its costs'*

The purpose in this section is to attempt to avoid failure by foresight and to increase the usefulness and robustness of a national value-added system. Whilst espousing the justifications for the development of performance indicators in terms of attempts to improve efficiency and equity, Smith identified a number of unintended consequences, a '*huge number of instances of unintended behavioural consequences of the publication of performance data'*. He named eight problems associated with non-effective or counter-productive systems:

- tunnel vision
- sub-optimisation
- myopia
- measure fixation
- misinterpretation
- misrepresentation
- gaming
- ossification

With the exception of ossification every one of these possibilities was commented upon by headteachers in open-ended items in the questionnaires. These are not theoretical problems but actual, already-perceived problems.

6.1 Tunnel vision

In education, tunnel vision might be manifested by the concentration on examination performance to the exclusion of other concerns. Equally the concentration on *academic* indicators and the neglect of *vocational* qualifications would be indicative of tunnel vision. Yet again, if only national curriculum subjects were included in value-added tables, this would lead to a concentration on only those subjects, neglecting many others that are also important.

At the same time as national value-added systems are launched, the need for a broader framework of indicators should be consistently acknowledged. In particular, reference could be made to the development

of on-going research on educational indicators with a broad range of concerns, within the European community. (Climaco, 1995; Emin, 1995; OECD, 1995; Scheerens, 1995)

6.2 Sub-Optimisation

Sub-optimisation is defined as '*narrow local objectives by managers, at the expense of the objectives of the organisation as a whole*'. At various levels in education this problem can be detected. With local management of schools there is an incentive to remove difficult pupils and concentrate on those for whom the measured performance will give credit to the school. There is also an incentive to put pupils in for easy examinations and to use selective entry policies whether this is good for the community in which the school functions or not. Sub-optimisation will be a problem for local education authorities trying to look after the interests of all pupils including difficult sub-groups of pupils in the face of schools that are intent on local objectives that will be publicly visible.

Smith also comments '*most outputs in the public sector are the result of team rather than individual efforts*'. As a result, if the implicit reward scheme is directed at individuals, sub-optimisation can arise. If too much emphasis is placed upon the supposed value-added contribution from a single individual (such as a single teacher or the headteacher), then the contribution that others make to this performance may fail to be recognised.

It would be unfortunate if the welcome that schools have extended to value-added turned to hostility due to the over-use of these indicators, in particular their use in staff appraisal. Those who have studied indicator systems urge that a broad range of indicators is needed in order to keep the whole range of outcomes and concerns of the system under review (e.g. Murnane, 1987). Nowhere would this be more important than in relation to staff appraisal. However, if used positively to give recognition to a few outstanding teachers, the impact could be more acceptable.

6.3 Myopia

If the focus is on immediately measured examination performance, then concerns regarding the suitability of various courses for subsequent study or employment may take lower precedence than the likelihood of the pupil showing a good value-added score in a particular subject. Yet, at a national level, it is the long-term benefits that are sought. Many innovations in education are justified in terms of creating better employees, more rounded adults, more conscientious citizens, reducing delinquency, etc. None of these important aims can be evaluated in the absence of long-loop feedback on long term consequences of various methods of schooling. The FEFC, the Careers Service and the Youth Cohort Study have all attempted to collect pupil destinations following departure from the education system. These disparate efforts need to be developed into a system to provide long term feedback to local authorities, schools and

colleges, with a view, ultimately, to some accountability for the long-term outcomes of local and national policies.

6.4 Measure Fixation

Defined as '*an emphasis on measures of success rather than the underlying objective*', a typical example in education is the impact of the reporting of the percentage of pupils attaining five grades in the range A* to C. This has led to a concentration on pupils likely to reach a 'D', with the aim of altering that particular and arbitrary indicator, rather than on the more acceptable objective of improving the standards of all pupils.

Particularly in social science, it is important to recognise that there are rarely single measures that produce single answers, even to simple questions. For example, the use of a curriculum-free aptitude test as an input can give additional information regarding pupils' progress; it does not yield the same indicators as a prior attainment measure. Some pupils may not be making the effort needed to do justice to their aptitudes, and a curriculum-free aptitude test can provide an indication of which pupils fall into this type of 'under-achievement' category.

If, instead of using *input* data, the performance of pupils in one subject is compared with their *concurrent* performance in other subjects, this also provides information of considerable interest. To make such comparisons accurately the differences in subject difficulties need to be taken into account, as in the Relative Ratings indicators that have been provided to schools in Scotland since 1991, and were provided in the *Review of 16 -19 Qualifications* (Dearing, 1995).

In short, it is undesirable to fix on a single indicator; more information can be obtained from more indicators and this can help to avoid the search for quick-fixes of the indicator rather than the long haul of real effectiveness.

6.5 Misinterpretation

> 'Many of the production processes in the public sector are immensely complex, and building a realistic model of them is severely taxing ... full account must therefore be taken of the external environment in which public sector organisations are operating ...'

Included in the environment in which examination results are produced is the matter of resources. If additional information is to be collected for Stage 2 analyses, resources should be included.

A widespread misinterpretation of value-added indicators is the confusion of correlation with causation. Strong associations do not prove that one element causes the other. For example, large classes are often found to be producing better value-added results than small classes. Can this be interpreted in a causal fashion - - -that large classes *cause* high achievement? There is a strong competing hypothesis which is that biddable, hard working pupils are placed in larger classes than difficult pupils and that alone could explain the positive correlation between progress and class size. Only the kind of experiments known as randomised controlled trials can provide strong evidence as to 'what works.' This, indeed, is the message of 'evidence-based medicine'. A UK

leader in the field, Professor David Sackett of Oxford's Centre for Evidence-Based Medicine, stated bluntly:

> "If you are scanning an article..... and it is not a randomised controlled trial, why on earth are you wasting your time"

If taken seriously this admonition might cut down the reading time needed to keep up with educational research to close to zero, unfortunately. The damaging effect of the widespread confusion of correlation with causation is to delay the efforts needed to obtain high quality data from controlled trials: evidence-based education.

6.6 Misrepresentation

This comes close to the 'fiddling' mentioned at the meeting with statisticians and also reflects Deming's warning: "*Wherever there is fear we get the wrong figures*." Smith suggested that misrepresentation can take two forms: 'creative' reporting and 'fraud'.

Creative reporting could well be applied to some truancy figures. Whilst it might seem improper to suggest such events, the experience of performance monitoring is that creative reporting occurs. Creative reporting of data is more likely to occur in difficult to define areas such as truanting rather than in measures of relative progress (i.e. value-added data).

Unintentional misrepresentation

Less culpable misrepresentations but ones that are probably more widespread and damaging, derive from misunderstandings. For example, reporting data to several decimal places is a mild form of misrepresentation, implying greater accuracy than is reasonable. Equating examination scores (eg between subjects) without supporting evidence that they can be considered to have the same 'tariff' value, is another misrepresentation.

As the widespread availability of value-added indicators comes closer, the pressure on external marking agencies and examination boards will increase. It is essential that adequate procedures are adopted to provide the public with answers to the kinds of questions that they might properly ask, such as comparability between Boards and syllabuses. Only by making such data regularly available and *non-consequential in its impact* on school value-added indicators (for example by the syllabus by syllabus analyses suggested in Section 5) can misrepresentation be avoided.

Whilst Examination boards seem to have been reasonably comparable as long as we were concerned only with fairness *to individual pupils*, this now becomes a matter of such importance *to schools* as to need formal and yearly attention. Schools and syllabuses are confounded in the datasets and this confound must be taken into account in the design of the system. Discrepancies between Boards and syllabuses that are such as to affect school's value-added scores, should be reported each year as a matter of course.

Quality Assurance procedures need to be adopted that include designing the examination and Key Stage systems with the following safeguards:

- Candidates' names are not on the examination scripts when these are marked
- School/examination centre names are not on the examination scripts and cannot be deduced from identification numbers. (The use of bar codes should be considered)
- Schools/examination centres should not send their examination scripts to a known marker
- Examiners and markers should not assess scripts from their own school
- There should be legal guarantees that marking is not subcontracted
- Standard practices (eg the use of check digits) for eliminating errors in identification numbers should be adopted.

Standards over time

Another source of potential misinterpretation based on misunderstanding is the statements made about standards over time. A value-added system each year provides comparative data: the relative progress of pupils compared to similar pupils in other schools. It cannot automatically be used to monitor standards of performance. The national regression lines each year will have approximately half the pupils above and half the pupils below the regression line.

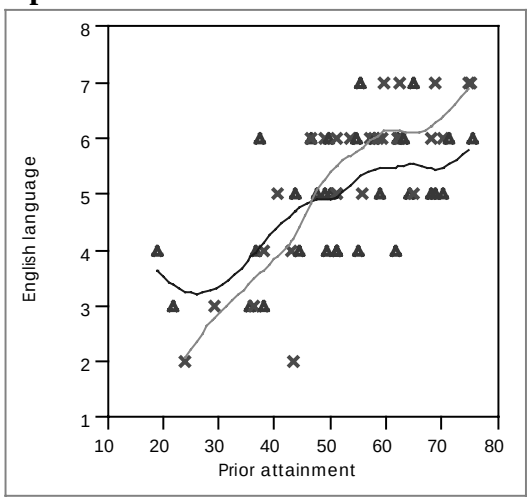
Standards over time are of interest in that, should the use of value-added feedback improve the performance of the entire system, we would not know unless a system was in place that enabled us to make some estimate of just how successful this mass attempt at accurate feedback had been. It would seem to be desirable, therefore, to establish an agency with an agreed methodology to address this complex issue.

Stereotyping and excuses

The issue of the use of additional variables is difficult and relates to misrepresentation of the unintentional kind. Taking sex into account makes a difference to value-added scores for some schools. Nevertheless, in teaching it should be *individuals* who are considered, not gender groups, ethnic groups, well-dressed groups or tall groups. Correlations of attainment with the latter two groups would possibly be present but should they become part of a reporting system? Should gender? Should not each gender compete on a level playing-field, as they will for jobs or university places? Does then the inclusion of sex in a value-added system misrepresent and confuse? Consider Figure 6.1, showing achievement in English Language among boys (blue triangles) and girls (red 'x's.) Note that the trend-lines are close together for the majority of pupils. and that scores are so intermingled that the only response to 'girls are doing better than boys' should be 'which girls? which boys?'. Differences in averages hide the fact of considerable overlap. If there is racism or sexism in marking, this should be prevented by the removal of names from papers. If there is racism or sexism in teaching this serious charge must rest on more than a statistical difference that could be due to many factors, such as developmental differences or friendship groups and social pressures in a particular school and beyond the responsibility of the school.

The important point is that we must be careful not to build stereotyping into statistics.

Figure 6.1 Achievements of boys and girls in GCSE English Language, with splines fitted to show non-linear trends



6.7 Gaming

> 'While misrepresentation leads to distortions in reported behaviour, gaming is the equivalent manipulation of actual behaviour. It is therefore potentially severely dysfunctional ...'

The instance of gaming that was commented on by many headteachers was the depression of an intake value in order to make the subsequent value-added scores look better. In primary schools this would involve being as severe as possible in the baseline assessment as pupils arrive in school. If baseline tests used teacher-given check lists rather than actual pupil responses, the opportunities for gaming would be considerable. The temptation to lean towards a severe interpretation would be strong. In secondary schools gaming would be a temptation particularly in Key Stage 4 where the output is the highly visible examination results at GCSE. Schools might feel they could sacrifice their performance at the end of Key Stage 3 for a good value-added score on the all-important terminal leaving examinations, GCSEs.

Clearly quality assurance procedures in test administration and marking, and complete standardisation, are the only way to ensure that the depression of baseline or intake tests is not used as a method of distorting value-added indicators.

Another opportunity for gaming is provided by the use of teachers' predictions of levels of performance. Indeed, SCAA's scrutiny picked up a glaring example of this during the time of this project.

> "The process of adjusting candidates' marks and grades without reference to the scripts, *so that these were in line with centres' predictions*, was unacceptable."
>
> TES 10-1-97, quoting from a SCAA scrutiny (*emphasis added*)

Whilst it would be attractive to think these strategies would have no appeal if the value-added information was only to be used internally for management information within the school, unless schools were guaranteed not to have to produce the data, there will be an incentive to make the data look as good as possible. Only confidential systems ensure that there is not only no incentive to corrupt the data but every incentive to

play the game as fairly as possible in order to know the situation of one's own school in comparison with other schools as accurately as possible. In an era of "freedom of information", confidential systems are not likely to remain confidential

Another source of gaming that is of concern to schools is the search for 'easy' syllabuses or subjects. For example, pupils planning to study Economics at university might choose to do English A level and thereby be likely to get higher grades than had they chosen Mathematics at A level. However such a choice might subsequently be a source of regret if difficulties were encountered on the Economics course in the quantitative elements. Staff report feeling under pressure to find easy syllabuses within subjects. Both these gaming strategies can be made irrelevant by publishing the regression equations each year for every subject and syllabus, as recommended in Annex F.

6.8 Ossification

Smith defines this as

> 'Organisational paralysis brought about by an excessively rigid system of performance evaluation.'

and suggests that

> '... any scheme is likely to be deficient to a greater or lesser extent, and so will need to be regularly reviewed and updated.'

The antidote to ossification is to have constant monitoring and revision of the system. Indeed Smith talks about 'messy and political, responding to new circumstances, the nature of which - almost by definition - is likely to be unforeseeable'. This recognition of unpredictability in a changing dynamic situation must apply to any monitoring system of social behaviour. It is taken care of in our recommendations in Section 7 by the setting up of a monitoring team representing various stake-holders and by the design of having multiple independent analyses undertaken to locate features in the data that might be useful in improving schools or in improving the monitoring system.

In conclusion, the possibility that publication of a few indicators can distort practice in undesirable ways is widely recognised. One response is to urge that no indicators be published. A better response is to urge that indicators are selected with extreme care and with particular attention to their impact on school practices. Further, it is highly desirable to use of a wide range of indicators, a practice strongly encouraged by the Further Education Funding Council.

6.9 The role of a national authority

Qualifications are an important kind of currency and value-added indicators are 'qualifications' for schools. There must be confidence in all qualifications and clarity as to their import. Setting up the necessary conditions for this to happen is the role of a national authority. Employers hire people with qualifications from around the country and need a system in which exchange values are known.

The role of a national authority must be to establish the standards, as in weights and measures or the value of the pound. This could not be left to individual providers: the framework of information that enables intelligent choices to be made, and that enables schools to learn accurately about their effectiveness, must be a national framework. Annex F presents the ‘fifty percent framework’, a way in which choice, diversity and standards can all be maintained.

SECTION 7: RECOMMENDATIONS

These recommendations address the two purposes identified for value-added work

- for internal school information
- for external accountability, including the possibility of publication

Recommendations are made under the following headings

7.1 Features of a national value-added system

1.1 Recommendation:

For internal school use, an initial value-added system should be developed as soon as possible, throughout the primary and secondary sectors, based on individual pupil data arising from national curriculum testing and public examinations. [Sections 2, 3, 4 and 5]

1.2 Recommendation: .

For internal school use, the value-added information should be presented to schools in forms that are readily understandable. [Section 2]

1.3 Recommendation:

For internal school use, value-added feedback should be provided economically and promptly as part of the national curriculum testing and examination processes, and concurrent with the reporting of test and examination results. This is referred to as 'Stage 1' [Section 1.2; Section 5.1]

1.4 Recommendation: .

For internal school use, the use of simple residual gain analysis on pupil level data, organised by syllabus/subject, should be adopted for initial feedback (Stage 1). [Section 1.2, 1.3; Section 2.2, Section 3, Annexes C and D]

1.5 Recommendation:

Consideration should be given to the voluntary inclusion of value-added profiles in the School Performance Tables or the publication of the data for those schools in the top 10 percent of schools for value-added in each major subject, but only for schools meeting the minimum cohort size requirements as per recommendation 1.6. [Section 2.4; Sections 4.7; 4.8; 4.9]

1.6 Recommendation:

Before there is use of value-added data for public accountability, a second stage of data analysis should be undertaken, using a variety of statistical techniques,

with particular attention paid to atypical schools. [Section 3.2, 3.3]. This **second stage of data analysis** (requiring about 3 to 6 months) should provide

- the **minimum cohort size** required for the calculation of reasonably reliable value-added indicators for the data that year; [Section 3.1; Annex E]
- simple and **adjusted value-added estimates for the few atypical schools** that are heavily influenced by factors such as sex composition and English second language needs; [Section 3.3]
- further exploration of the data with the purpose of supporting **school improvement.[**Section 6]

1.7 Recommendation:

Before there is use of value-added data for public accountability, consideration should be given to methods of taking into account the considerable instability in the value-added indicators from year to year. Publication should be based on at least three years' data.

1.8 Recommendation:

Before there is use of value-added data for public accountability, consideration should be given to using a value-added profile to represent the work of schools, rather than a single indicator. [Section 3; Sections 4.7; 4.8; 4.9; Section 6]

1.9 Recommendation:

Before there is use of value-added data for public accountability, further work should be undertaken to arrive at acceptable and robust methods of recording and taking into account pupil attendance and mobility, including guidelines for allowing the non-reporting of data for some pupils for reasons associated with attendance or other complicating factors. [Section 3.3, Finding 13]

7.2 Infra-structure: technology and training

2.1 Recommendation:

Provision should be made for extensive INSET on value-added for headteachers, teachers, governors and parents. Consideration should be given as to whether there should be centrally commissioned production of materials, including software, or funds made available to schools to purchase from independent providers working to specifications.[Section 2]

2.2 Recommendation:

The qualifications for headteachers should include a component on the interpretation of value-added data [Section 2; Section 4]

2.3 Recommendation:

The curriculum for Initial Teacher Training should include sufficient statistical training to form the basis for a good understanding of value-added analyses. [Section 4]

2.4 Recommendation:

Guidance should be issued to schools regarding the retention of Key Stage and GCSE data for use in value-added systems. Consideration should be given to whether this guidance should be underpinned by legislation.[Section 5]

2.5 Recommendation:

A unique pupil identifier, for use in keeping track of children during the years of compulsory schooling, should be introduced with the use of check-digits or other methods of ensuring accuracy.[Section 3.4; Section 5]

2.6 Recommendation:

Specifications should be developed for computer-based management information systems to be used in schools, including data-checking on entry, ease of importing and exporting and electronic data interchange capacities to facilitate national data collection. [Section 3.4; Section 5]

7.3 External Marking Agencies and Examination Boards

3.1 Recommendation:

In view of the accountability implications of value-added analyses, it is necessary to introduce more stringent quality assurance procedures into the marking and grading of end of key stage tests and public examinations. Consultation should begin on a national system of reporting, following every test or examination:

- internal consistency measures
- inter marker reliability
- examination result consistency across syllabuses [Section 6.6]

3.2 Recommendation:

Examination Boards, and others in a similar role, should be given not longer than the academic cycle of 1998-1999 to implement the quality assurance procedures that will be essential in view of the accountability aspects of value-added work. [Section 5, Section 6]

3.3 Recommendation:

Subsequent to the initial stage 1 feedback to schools, a national agency or agencies, or designated existing organisations, should

- archive data securely, with confidentiality safeguards and meeting the requirements of the Data Protection Act;
- at the request of the Assessment Authority conduct monitoring studies of the test / examination process;
- report each year subject / syllabus regression lines.[Section 3; Section 5]

7.4 Primary Sector Recommendations

4.1 Recommendation:

In the primary sector the development of baseline tests and their adoption by schools should continue to be encouraged.

4.2 Recommendation:

KS1 testing should eventually be looked at in the context of intake levels on baseline assessments. That is, baseline-to-the-end-of-KS1 should form a value-added system.

4.3 Recommendation: .

The tests being developed by SCAA for the end of Year 4 should be evaluated for its suitability for the provision of two further primary school value-added systems after the baseline-to-KS1 system.

4.4 Recommendation:

For the years 1997 and 1998 value-added analyses for Key Stage 2 should be provided to primary schools where possible, for internal use only, with the data belonging to the schools. Because not all schools have Key Stage 1 data from 1993 and 1994, these analyses should be carried out by commissioned agencies with a view to collecting the maximum amount of available data. The tender should include requirements for the provision of supported feedback to schools with data and the development of systems for recording mobility and attendance.

7.5 The Primary-Secondary Transitions

5.1 Recommendation:

Consideration should be given to methods for the transfer of data on achievement, attendance and mobility from primary to subsequent schools

5.2 Recommendation:

Special consideration should be given to value-added systems that support feedback for middle schools and other schools that enrol pupils part way through a key stage.

7.6 Secondary Schools

6.1 Recommendation:

In 1997 regression lines for KS3 to GCSE results should be developed and made available to schools, in a form suitable for use in spreadsheets or statistical packages. This exercise should be used to prepare the way for the introduction of a full national system in 1998.

6.2 Recommendation:

From 1998, schools should be provided with value-added scores based on KS2 data predicting KS3 outcomes, but this data should not be *published* until three years of such data has been collected in the context of adequate provision for the concerns indicated in recommendation 1.5. [Section 3.1, Finding 6]

7.7 Monitoring The Value Added System

7.1 Recommendation:

Representatives of various stakeholder groups should form a consultative committee to monitor and advise on the on-going development of the national value-added system. Teachers' associations, LEAs and similar organisations, the inspectorates, governors, parents and researchers, (including researchers from other countries seeking to implement similar systems) should be enabled to monitor the implementation and impact of value-added systems [Section 6]

7.2 Recommendation: .

Methods for monitoring standards over time should continue to be developed for the national testing arrangements, as this issue is not addressed by value-added yearly figures.

7.8 Extending the benefits of value-added systems

8.1 Recommendation:

Methods of reporting value-added for pupils with Special Educational Needs should be the subject of research and development .

8.2 Recommendation:

The value-added framework should be extended to include vocational and other qualifications (Annex F).

8.3 Recommendation:

At the same time as national value-added systems are launched for end of key stage tests and examination data, the need for a broader framework of indicators should be consistently acknowledged.

8.4 Recommendation:

Schools should be encouraged to monitor a broad range of indicators as part of a regular cycle of self-evaluation. .

8.5 Recommendation:

Consultations should start on the possible use of value-added indicators as an outcome measure for LEAs.

8.6 Recommendation:

In order to learn the long term impact of the value added by schooling (and, therefore, in order to evaluate the cost-effectiveness of schooling practices) indicator systems should be developed for long-term outcomes.

8.7 Recommendation: .

In the longer term, projects employing experimental methods should be funded in order to encourage evidence-based professional practice.

ANNEX A: LIST OF NINE REPORTS

Technical Reports Relating To Primary Schools:

Tymms, P. B., & Henderson, B. (1995). *The Value Added National Project: Primary Technical Report* ***(Primary Technical Report No. 1)*** ISBN 1 85838 094 4. School Curriculum and Assessment Authority, London.

Tymms, P. B. (1996a). *The Value Added National Project: An analysis of the 1991 Key Stage 1 assessment data linked to 1995 KS2 data provided by Avon LEA* ***(Primary Technical Report No. 2)***. London: School Curriculum and Assessment Authority.

Tymms, P. B. (1996). T*he responses of headteachers in Avon to value-added feedback* ***(Primary Technical Report No. 3).*** London: School Curriculum and Assessment Authority

Tymms, P. B. (1996b). *The Value Added National Project: Value Added Key Stage 1 to Key Stage 2:* ***(Primary Technical Report No. 4)***. London: School Curriculum and Assessment Authority.

Technical Reports Relating To Secondary Schools

Trower, P., & Vincent, L. (1995). *Value Added National Project: Secondary Technical Report* ***(Secondary Technical Report No. 1)*** ISBN 1 85838 095 2. London: School Curriculum and Assessment Authority.

Trower, P., & Vincent, L. (1996). *Value Added National Project: Experimental study of different kinds of data presentation* ***(Secondary Technical Report No. 2).*** London: School Curriculum and Assessment Authority

Vincent, L. (1996). *The Value Added National Project: analysis of KS3 data for 1994 matched to GCSE data for 1996.* ***(Secondary Technical Report No. 3)***. London: School Curriculum and Assessment Authority

General Reports

Fitz-Gibbon, C. T. (1995). *Issues to be considered in the design of a national Value Added system* ***(Interim Report)*** ISBN 1 85838 084 7 London: School Curriculum and Assessment Authority.

Fitz-Gibbon, C. T. (1996). *Feasibility studies for a national system of value-added indicators* ***(FINAL Report).*** London: School Curriculum and Assessment Authority. ISBN 1 85838 249 1

Available from SCAA Publications, PO Box 235, Hayes, Middlesex, UB3 1HF and QCA website

ANNEX B The VANP Advisory Group

Phillip Evans	Authority Member
John Marks	Authority Member
John Gardner	Chief Statistician, DFEE
Peter Matthews HMI	OFSTED
Colin Robinson	Head of Research and Statistics, SCAA
Robert Wood	Curriculum and Assessment Division, DFEE
Barry Creasy	Research and Statistics Unit, SCAA
Phase 1 only:	
Nick Tate	Chief Executive, SCAA
Bill Scott	Assistant Chief Executive, SCAA
Jean Haigh	SCAA Primary team
Phase 2 only:	
David Hawker	Assistant Chief Executive, SCAA
Alison Jeffrey	Curriculum and Assessment Division, DfEE
Mike Kilyon	Bradford LEA

From the CEM Centre, University of Durham:

Carol Fitz-Gibbon	Project Director
Peter Tymms	Director, Primary
Paul Trower	Statistician
Brian Henderson	Research Associate
Luke Vincent	Research Associate

Other participants:

Audrey Brown	DfEE
Chris Bryant	OFSTED
Trevor Knight	DfEE

ANNEX C: MODELLING ISSUES

Some statisticians have strongly urged us to recommend, for the Stage 1 feedback, an analysis that uses multi-level models and provides a direct estimate of 'the school effect'. This annex presents an introduction to this apparently contentious issue. It would be possible simply to argue that the need for the complex multi-level modelling approach must be established by its advocates, especially since all the technical reports for this project have shown how little difference the use of multi-level modelling makes to value-added indicators. However, in the interests of the continuing debate, here is an introduction to the issues.

A minimal assumptions approach to data description: residual gain analysis

Throughout this report, and constantly in work with schools, the simple definition of value-added has been 'relative progress', the difference between a pupil's attainment statistically predicted from knowledge of the data, and the actual attainment. (Figure AC 1). The statistically predicted score can be roughly described as the average score obtained by similar pupils in other schools, so it provides a fair and understandable comparison for the score obtained by one's own pupil.[17] This difference is the progress made by one's own pupil compared with that made by other similar pupils: i.e. the relative progress. This conception is illustrated in Figure AC1.

More complex analyses

The argument made by some statisticians is that the school should be explicitly modelled, even in the first stage of feedback (Figure AC2). Such modelling can be undertaken in generally-available statistical packages such as Minitab or SPSS by use of a procedure for comparing means (e.g. analysis of variance, fixed or random effects) or the school can be modelled in a more sophisticated way by a procedure that takes account of the fact that pupils within a school are more alike than randomly allocated groups would be. This procedure is called multi-level modelling or hierarchical linear modelling. The multi-level modelling procedures 'draw strength' from the whole body of data when estimating school effects. Thus small schools will have their estimates moved towards the average for their intake; the smaller they are, and the more extreme the scores, the more they will be adjusted towards average. This procedure is illustrated in Figure AC3 (not to scale - the adjustments are generally very small).

Kreft, a statistician who has written extensively in the US on multi-level models, described the situation thus:

> *Unreliable estimates are shrunken upwards, or downwards, towards the average value calculated over all schools. Parameter estimates based on small numbers of observations have large sampling errors, resulting in more shrinkage than estimates based on large numbers of observation. This shrinkage to the mean is related to 'borrowing of strength" by small schools from large schools. But ranking such shrunken estimates means that well performing schools, but represented by a small number of observations, are shrunken downwards, and badly performing schools with*

---

[17] A regression line is used rather than a spline or other smoothing technique, but this choice generally makes little difference and it keeps the procedure open and simple. The outcome measure is discrete rather than continuous, in any case, and the interpretations teachers make often allow for inherent unreliability in the grades awarded.

small numbers of observations, are shrunken upwards to the mean. As a result two widely different schools may end up close together in rank, appearing more alike than they are.

Kreft, I.G.G. Are Multilevel techniques necessary? An overview, including simulation studies. 1996 WWW

It has been suggested elsewhere that schools would prefer to see unshrunken data and that, since the school effect for a year is based on a population, 100 percent of pupils, not on a sample, shrinkage is not necessarily required and would have the following undesirable consequence:

Suppose a school serving students with low prior achievement were especially effective. In this case, it would have its score 'pulled' toward the expected value of schools with children having low prior achievement. That is, it would have its effectiveness score "pulled" downwards, in the "socially" expected direction, demonstrating a kind of statistical self-fulfilling prophecy! (On the other hand a similar school doing very badly would be pulled up.)

Raudenbush cited in Fitz-Gibbon 1991.

What would be the effect of using multi-level modelling for Stage 1 feedback?

To model 'the school ' is to make an assumption that there is some consistent 'effect' from the school on each pupil, either a constant effect, exactly the same for all pupils as in the equation in Figure AC1, or, as the modelling becomes more complex, some other effect working in the same way on all pupils. It is the view of this writer that such an assumption is unlikely to be valid, and until effects are traced to causes, the interpretation of the residuals should be left to the schools, not modelled, by multi-level modelling, ANOVA or in any other way. At the very least, there is virtue in this minimal approach *for the first stage of analysis*. In science, a declaration of ignorance is often more productive than an ill-supported assumption.

School or classroom

Finally, a very important point is that **'the school' is not the correct second level in the multi-level modelling. It is the teaching group that delivers instruction and should be modelled.** Information on teaching groups would not be acceptable as part of a national system, yet the 'school' effects will be different if the teaching group is or is not included as a second level (Secondary Technical Report no. 3).
Kreft remarks that

"If random coefficient models make a difference, it is not yet established where and under what conditions".

This certainly seems to be the case.

Figure AC 1: Value-added as the difference between an actual grade and the statistically predicted grade

The 'statistically predicted' score depends upon:

- the overall level of performance for that outcome, that year (the height of the regression line
- the relationship between intake and outcome measures that year (the slope of the regression line)
- the pupil's prior attainment

The equation showing the predicted score for individual student, i, is:

$$\text{Predicted } Y_i = \mu + \beta X_i \quad \text{---Equation 1}$$

In words, the predicted outcome is equal to the sum of

- a mean which indicates the measuring scale for the outcome (and will vary from year to year , from subject to subject as the overall level fluctuates slightly. The mean is simply a feature of the Y scale.

• a component due to the individual's prior attainment (X_i) and the slope of the regression line (β, beta)

The actual score will be represented, for pupil i , as

$$\text{Actual } Y_i = \mu + \beta X_i + e_i \quad \text{----Equation 2}$$

where e_i is the 'error' or 'residual' for a particular individual pupil

The difference between the predicted and actual is variously called the 'error', the residual or 'value-added' for that individual pupil. This is illustrated in Figure AC.1

Figure AC.1 The value-added residual as the difference between predicted and actual score

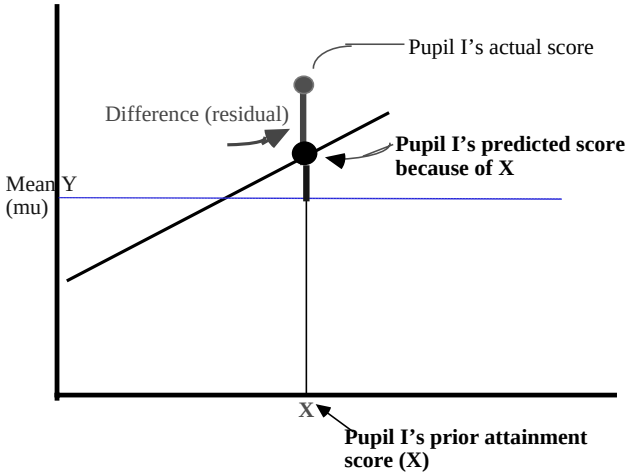


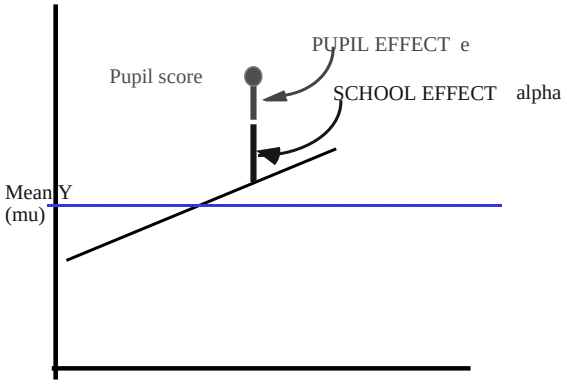
Figure AC2 Modelling the school

When the school is explicitly modelled the pupil's grade is broken into components, one attributed to the student and error and the other to the school. The equation becomes;

$$\text{PREDICTED } Y_i = \mu + \alpha_j + \beta X_i \quad \text{----Equation 3}$$

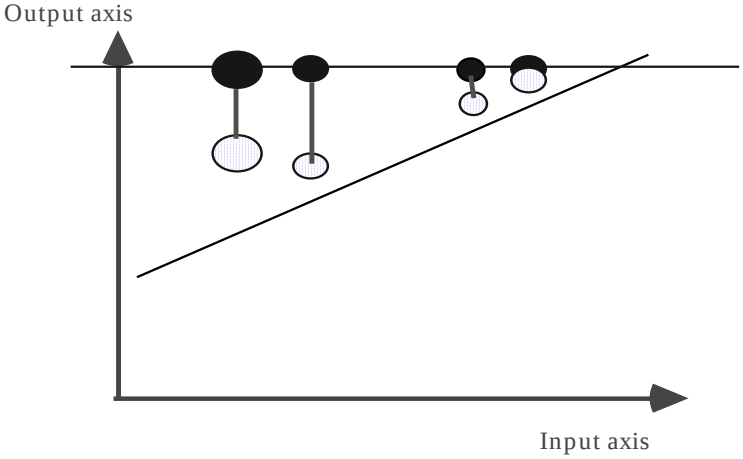
where α_j is the effect of the school

Figure AC.2 The value-added residual as the difference between predicted and actual score



The simple approach breaks down entirely if the school becomes an element in the model. If we move from residual gain analysis to a fixed or random effects ANOVA, or to multi-level modelling with the school at the second level, then the predicted grades are no longer simply the grades similar student obtained in the national dataset.

Figure AC3
A means-on-means graph to illustrate multi-level modelling's shrinkage



Key:

The actual, raw, mean value-added score for a large school

The shrinkage, to 'draw strength' from the whole dataset

The adjusted value-added score from multi-level modelling

The actual, raw, mean value-added score for a small school

The adjusted value-added score for the small school

ANNEX D: ERRORS IN MULTI-LEVEL AND OLS

Peter Tymms July , 1996

At the recent meeting at SCAA (Tues. 2nd July 1996) we were assured that one reason why multi-level modelling (MLM) was vital in the calculation of Value-added scores was that ordinary least squares (OLS) procedures produced misleading estimates of the errors on the schools based residuals. It was emphasised that ALIS in particular was misleading schools by suggesting that the errors providing errors that were smaller than they should be.

In what follows errors on Value-added scores for schools were calculated using OLS and MLM procedures. The Avon dataset, which includes some 247 schools and 7701 pupils with data on the 1991 KS1 results matched with the 1995 KS2 results, was used.

A very simple approach was taken in which KS1 data was used as a control for the KS2 average level across mathematics, English and science. Details are given below:

OLS

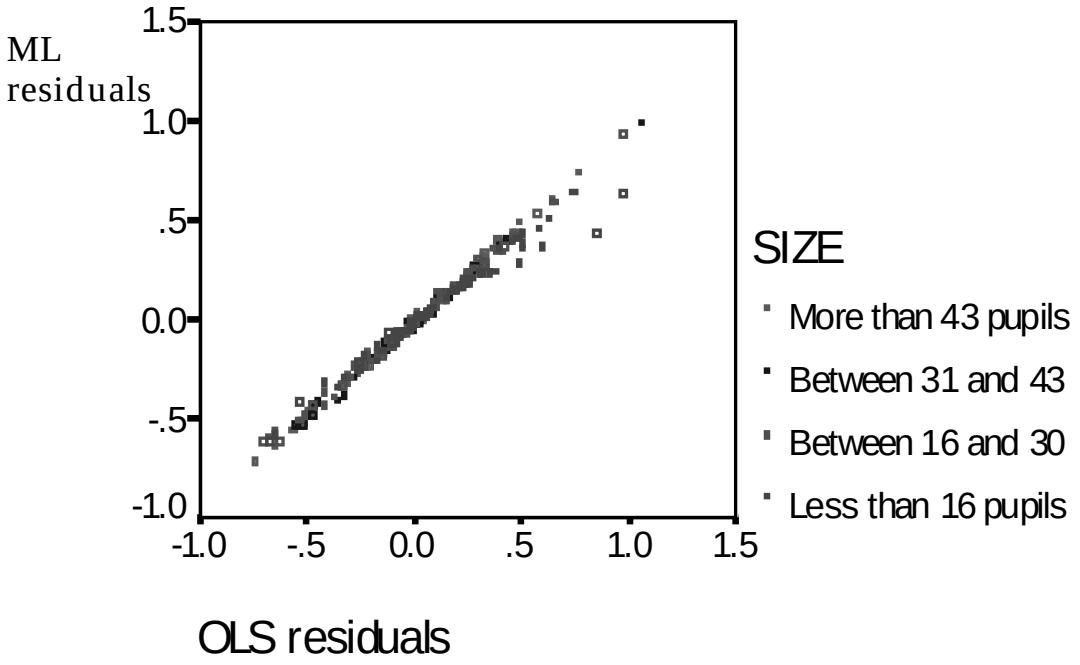
```
Multiple R .63485
R Square .40304
Adjusted R Square .40296
Standard Error .53994
------------------- Variables in the Equation ----------
--------
Variable B SE B Beta T Sig T
KS1 1.071279 .014871 .634851 72.036 .0000
(Constant) 1.653238 .028277 58.466 .0000
```

MLM

```
PARAMETER ESTIMATE S.  ERROR(U) PREV.  ESTIMATE
CONS 1.511 0.03403 1.51
KS1 1.159 0.0146 1.159
rand
LEV.  PARAMETER (NCONV) ESTIMATE S.  ERROR(U) PREV.  ESTIM CORR.
-------------------------------------------------------------------------------
 2 CONS /CONS ( 4) 0.09533 0.009438 0.09534 1
-------------------------------------------------------------------------------
 1 CONS /CONS ( 9) 0.1997 0.00327 0.1997
```

Unexpectedly the reported SE on the slope coefficient was greater in OLS than in the MLM although there was little in it. The error on the constant was slightly greater in MLM. As expected the graph in Figure AD.1 shows there was little difference between the OLS and MLM Value-added scores.

Figure AD.1 Errors in MLM and OLS, related to sample size



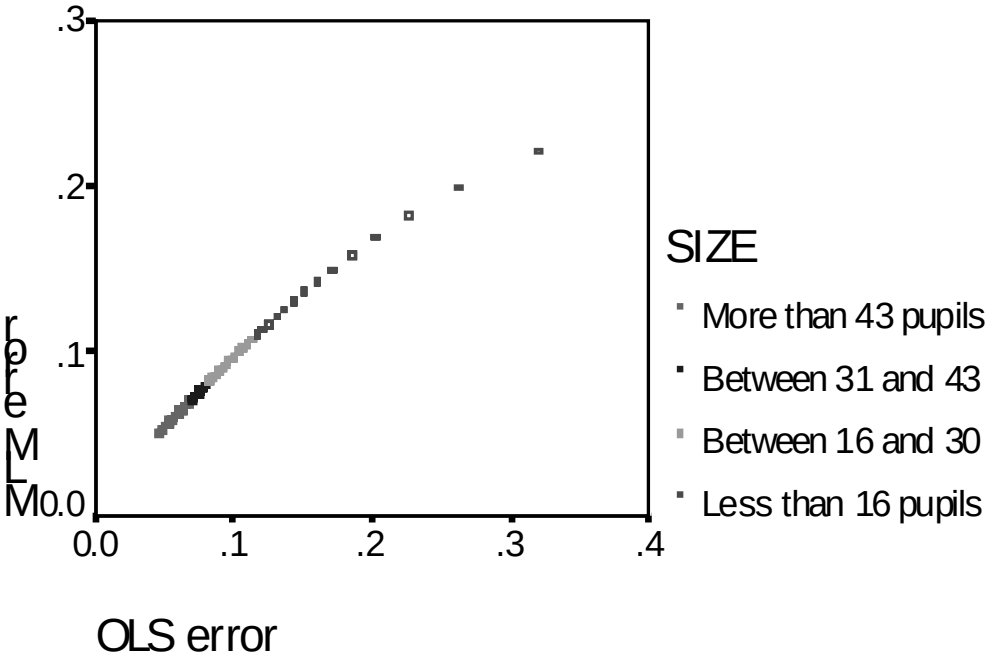
The diagram below shows the relationship between the MLM and OLS Standard Errors.

But the interesting question concerns the errors on the calculated school effects.

To do this in MLn the 'RESI' command was used with the setting 'RTYP 1' which is recommended when comparing units. The estimated variance of the residuals at the second level (the school level) were then square rooted to produce the MLM Standard Errors.

In OLS the pupils level residuals were used to calculated a "pooled SD". (A One-way ANOVA run in Minitab was used since this gives an immediate estimate of the pooled SD.) Then in each school the SE was calculated using the pooled SD divided by the square root of the number of pupils in that schools cohort.

Figure AD.2 Errors in MLM and OLS on value-added indicators



For most schools there was essentially no difference between the two sets of errors. For small schools the estimated errors were slightly greater for OLS.

This OLS procedure has been used for years within ALIS.

ANNEX E: RELIABLE VALUE ADDED MEASURES

When individuals are tested, it is considered important to have a test that provides a reliable measure, generally one with an internal consistency of about 0.9.
Analogously, it is important that if schools are to be 'tested' and assigned a value-added score, this is based on a reliable measure, especially if the information is to be publicly available. Winer (1971, p.285) or Goldstein (1987) can be consulted for the derivation of relevant formulae. These are applied here to investigate the issue of the size of sample needed, such questions as, for example:

- Do small primary schools contain a sufficient number of pupils to justify reporting their value-added scores?
- Will be the scores from secondary school departments that contain, each year, only a small number of pupils be sufficiently reliable?

The answer can be informed by locating the dataset on a diagram, using two pieces of information: rho and the range in sizes of the cohorts (See Figure E.1 below).

Rho

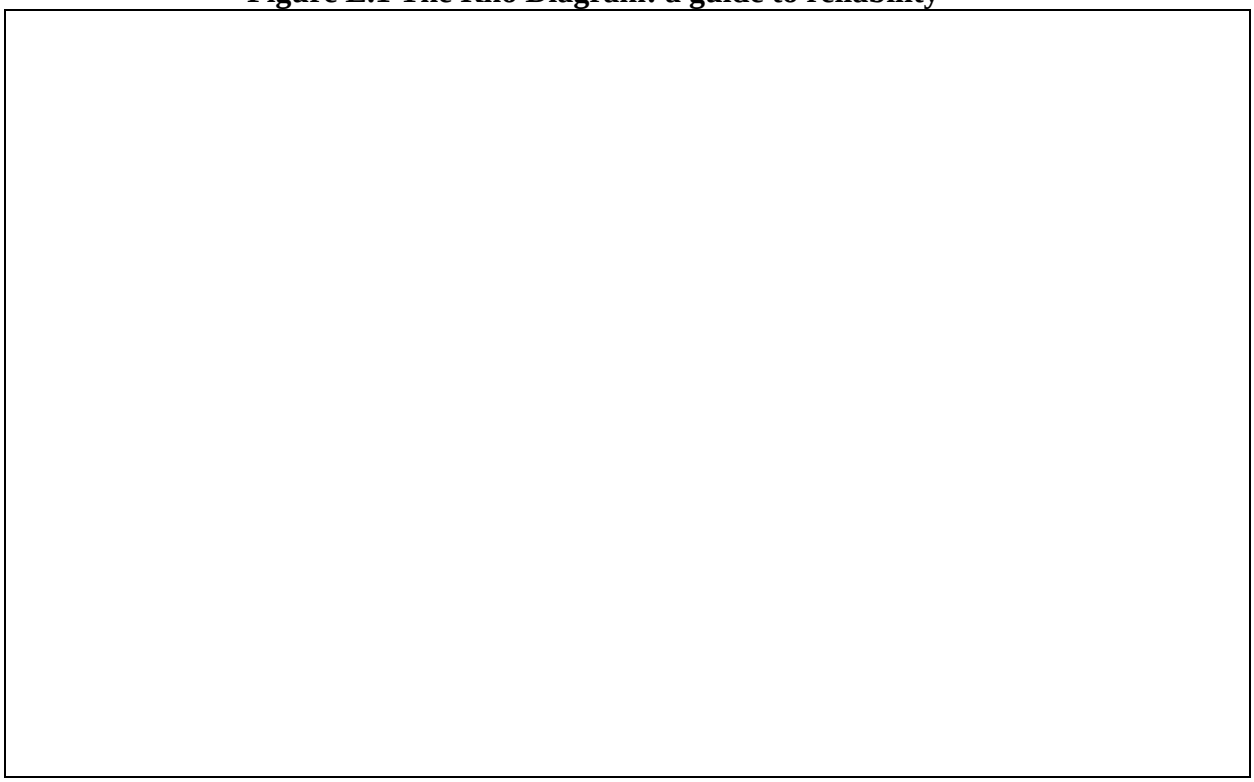
The quantity rho, which is explained further in the glossary, is a measure of the extent to which schools differ on a particular variable, such as on value-added scores. Rho is an index that will be different for each dataset, such as a set of primary schools, or a set of mathematics departments.
If rho is large it implies that there are differences *between* schools that are large when compared with the differences *within* a school. (This concept is illustrated by diagrams in the Glossary under the Rho entry.) It is the *relationship* between the variance *within a school* to that *between* schools that determines the extent to which schools look different in their putative effects on pupils.
Rho is analogous to r^2 the proportion of variance accounted for, as it represents the proportion of the total variance accounted for by schools.

The rho diagram

Once the value of the quantity called 'rho' is known, this can be located on the horizontal axis and then, by moving vertically to a point on the curved line, the minimum size of group can be read off the vertical axis. For example if rho was 0.1, then a group size of at least 90 pupils would be needed to reach a reliability of 0.9 for the value added indicator.
This kind of information enables us to look at current datasets and see if there is a chance of the value-added indicators being sufficiently reliable.

It would also allow a national system to set cut-off values for group sizes each year.

Figure E.1 The Rho Diagram: a guide to reliability



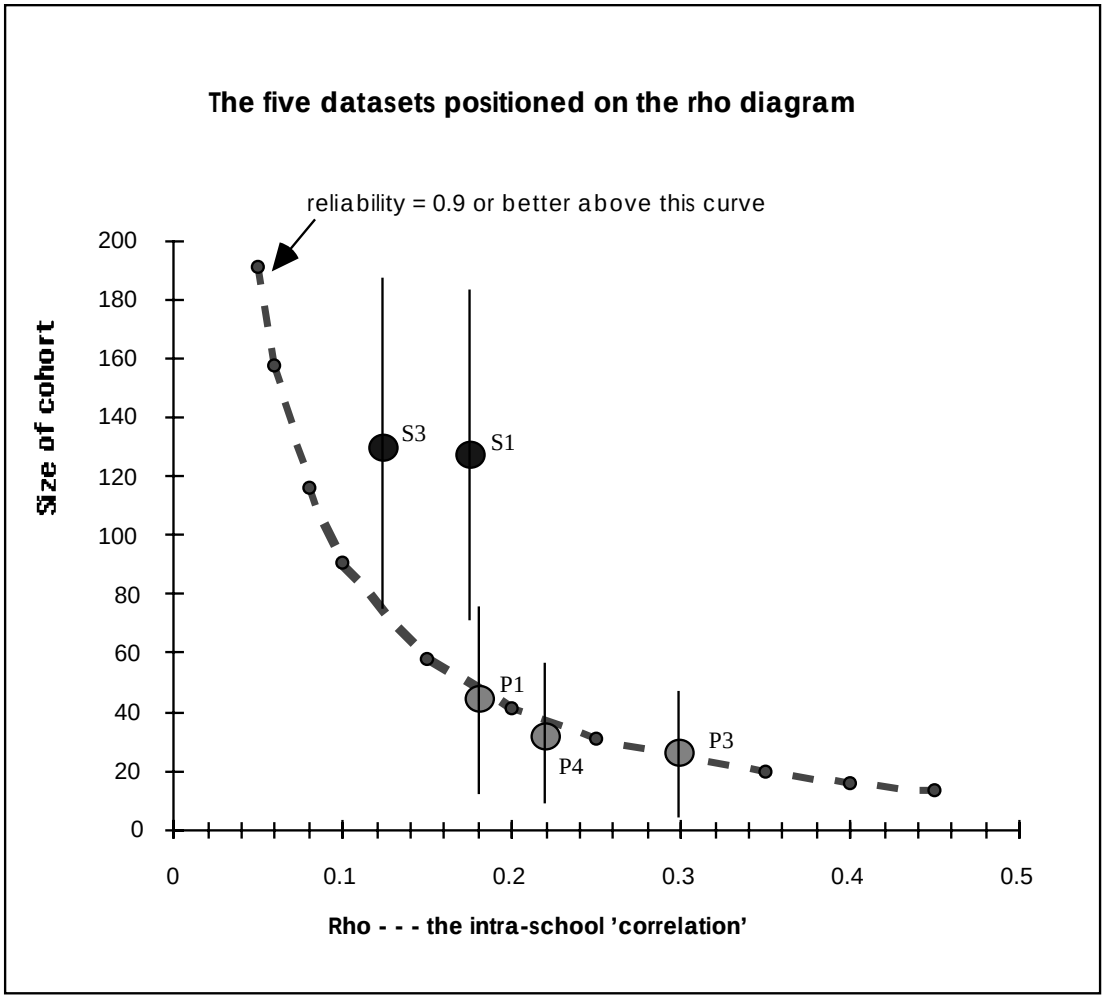
Applying the rho diagram to the datasets in Table 3.1

Values of rho for the samples investigated in the value-added National Project were reported in column 9 of Table 3.1. Rho for the first primary sample (P1) was 0.18 for an outcome measure consisting of an average score on end of Key Stage 2 tests . The average cohort size for this dataset was 45 (column 4 in Table 3.1). These two pieces of information locate dataset P1 on Figure E.2 as shown. It lies on the red curve showing that indicators for schools with about 33 pupils in the value-added cohort would have a reliability of 0.9. Smaller cohorts would have less reliable indicators. In general, only about half the primary schools in each of the three samples would appear to have had sufficient numbers of pupils to report reliable value-added indicators.

For the secondary samples, vertical lines have been drawn representing the spread of cohort sizes (eg a standard deviation of 58 for S3 in Table 3.1) and we can see that most schools would have adequate size samples.

Like all rules-of-thumb these procedures will be indicative rather than absolute but they represent the most informed approach and would also have the advantage of varying from year to year and group to group. This would discourage 'gaming' such as, for example, trying to avoid letting pupil numbers go above a pre-announced cut-off point.

Figure E.2. Samples Located On A Reliability 'Map'



S1 = Secondary '95
S3 = Secondary '96
P1 = Primary 1, 1995
P3 = Primary 3 1995 (Avon)
P4 = Primary 4, 1996

one standard deviation above and below mean cohort size
Located by rho and Mean cohort size,

The issue of the estimated reliability of the value-added indicator only becomes of great concern if indicators are to be published. In the use of value-added in schools, the evaluation of the success or otherwise of each pupil is undertaken with a knowledge of all the many contributing factors. Small numbers can be seen to be unstable, in that if one were deleted the average value-added figure would clearly change.

Policy implications

In summary, it can be recommended that for the purpose of publication the reliability of value-added indicators should be about 0.9 or better and this would rule out the publication of results for groups smaller than the size indicated by reference to 'the rho diagram', thus frequently ruling out publication for group sizes less than about 30 or 40, based on the data for one year. As several years' data are collected the average may become reliable enough to justify publication of, say, a three year rolling average. This will depend on the stability or consistency found from year to year, another kind of reliability, analogous to test-retest reliability in testing individuals.

The generally low value of rho (between 10 and 20 percent) should dampen enthusiasm for over-interpreting school effects. Much of the variation in results is *not* associated with the school attended.

An additional finding of interest: the impact of teaching group (class). Exceptionally high values of rho, however, were found when the group considered was not a whole year group but the teaching group (Secondary Report 3). The proportion of variance accounted for was 42 percent in mathematics and 33 percent in English classes, far more than the usual 10 to 20 percent typically reported in school effectiveness research. These high values might have been partially caused by the effects of streaming but could also have represented the impact of a teacher who will usually have taught the same group, alone, for two years. This was data from the last two years of compulsory secondary schools (Y10 and Y11) In other datasets the teacher-effect is hidden, by different teachers each year in primary schools, or several teachers often teaching a single group at A-level.

ANNEX F: THE FIFTY PERCENT FRAMEWORK

Extending value-added measures to all qualifications

A qualification represents an item of public information which is widely used by employers and by those charged with offering places in further and higher education. The qualification provides two signals:

- a certain content has been covered
- a certain level of accomplishment as been achieved in that content.

In other words a qualification signals **content-covered** and **level of performance.** Employers, admissions officers and others all need to be accurately informed about the content covered *and* the standards reached. Since the level of performance will depend to some extent on the developed abilities of the pupil, the qualification also indirectly signals **levels of developed abilities.**

The content covered requires qualitative judgements and those kinds of investigations of what 'stakeholders' expect that have been conducted by staff of the National Council for Vocational Qualifications under their Fitness for Purpose studies (e.g. Coles & Matthews, 1996; Coles, 1996). This kind of standard-setting, one responsive to the perceived needs of employers and higher education , must be on-going and is not absolute. Fitness for purpose in the 1990s was not the same as in the 1960s. Syllabuses and methods must also change. This need for rapid and responsive change is a major reason why choice and diversity should be provided throughout the system.

No single examination or assessment system should hold a monopoly. Not only would this lessen responsiveness but restrictions in syllabuses would undermine teachers' innovation, morale and their need to be responsive to the changing demands of their disciplines. **Such choice and diversity in content can be built into the system whilst nevertheless having convincingly comparable standards, at least as comparable as those that currently pertain in GCSE and A-level examining. A variety of syllabuses should be available and should be chosen on the basis of their content, suitability for purpose and their motivational properties for teachers and students.**

By implementing value-added assessments syllabus by syllabus, (as recommended in this report) there will be no advantage for schools in 'shopping around' for 'easy' syllabuses. Schools will be judged by the relative progress of their pupils on the syllabuses chosen. The 'difficulties' will be monitored and reported each year. There is no need to make all examinations equally 'difficult' at any one stage nor to restrict the knowledge base made available to pupils at any age.