Feedback, Value Added and Teachers' Attitudes: Models, Theories and Experiments

Thesis submitted for the degree of PhD,

by

Robert Coe

School of Education, University of Durham,

July 1998

ABSTRACT

A review of the literature shows a long history of research on the effects of feedback on performance. Feedback is found to enhance performance when it focuses attention on, or increases the saliency of, desired outcomes, or when the information it conveys helps to diagnose shortcomings in performance.

A critical review of school effectiveness research draws attention to the lack of existing evidence about the effects of attempts to improve students' academic performance. The lack of a sound theoretical understanding of the mechanisms by which schools and teachers may influence achievement is also discussed.

Two experiments were performed. In the first, 44 teachers of A level in volunteer institutions completed questionnaires designed to elicit their uses of and attitudes towards feedback and their self-perceptions. Teachers were randomly allocated to feedback or control groups, the former receiving information about their own students' value added performance and attitudes in previous years. The same questionnaire was used before and after distributing the feedback. Qualitative data were also collected and analysed. In the second experiment, a random sample of 192 institutions was drawn from the membership of ALIS and departments were allocated to one of three feedback 'treatments' or a control group. In each experiment, student examination performance before and after the intervention was compared.

In the first experiment, some attitude changes for the teachers were found between pre- and post-test, but the validity of the constructs measured by the questionnaire was somewhat challenged by the evidence from the interviews. Student A level performance for those whose teachers received the feedback was about a third of a grade better than for those in the control group, after statistical adjustment (effect sizes from 0.2 to 0.3).

No significant differences were found between any of the treatments in the second experiment.

For Jo,

without whose help and support
this would not have been possible.

TABLE OF CHAPTERS

DETAILED TABLE OF CONTENTS

LIST OF FIGURES, TABLES AND APPENDICES

Figures

Tables

Appendices

ACKNOWLEDGEMENTS

I owe a huge debt to my supervisor, Carol Fitz-Gibbon, for advice, encouragement and criticism in the production of this thesis, and I am truly grateful for her more than generous help. I wish to thank Joanne Frampton, Peter Tymms and David Galloway for comments on earlier drafts of parts of the thesis. I should also like to acknowledge the support of the ESRC in funding my fees, maintenance and expenses.

No part of this thesis has previously been submitted for any degree. An article based on Chapter 3 entitled 'Can Feedback Improve Teaching?' has been published in *Research Papers in Education* (Spring 1998: Vol. 13, No. 1, pp. 43-66), and its final form in that journal has benefited from the comments of the anonymous reviewers. A further article, derived in part from the content of Chapter 2 and jointly authored by myself and Carol Fitz-Gibbon, entitled 'School Effectiveness Research: Criticisms and Recommendations' has recently been accepted for publication by the *Oxford Review of Education* (to appear in December, 1998). The content of this thesis, however, while it owes much to the wisdom, advice and comments of those who have influenced my thinking, is entirely my own responsibility.

Robert Coe

Chapter 1

Background to the Study

Origins of the enquiry

The seeds of this research came from my own experience of receiving and analysing value added performance feedback while teaching. I was motivated mainly by curiosity to look at the results of the students in my department, and to see if there were any patterns or interesting features in the data. However, as I proceeded with the analysis – and in particular the analysis of individual teaching groups – I realised that I was producing something that might be interpreted as a measure of the performance of the teacher. Clearly, there were issues about the validity of such an interpretation, but my main feeling was that *my* performance was being judged, and that the act of judgement somehow seemed to make it more important to be seen to be doing well. I wondered whether other people would respond in the same way, and, if so, whether giving teachers this kind of individual performance feedback would lead generally to improved performance. Alternatively, it seemed possible that providing such feedback would have little impact on well established patterns of behaviour or objectives – much less on measurable outcomes – or that, even if it did, it might contribute more to an increase in anxiety than to a genuine improvement.

My interest in the general benefits of feedback was perhaps of even longer standing. My experience of teaching had been that when students were able to get good quality feedback about their progress they seemed to have a more positive attitude towards their work and a much better awareness of what they had to do. I have since learnt that the use of this kind of 'formative assessment' has been more

widely found to produce learning gains 'amongst the largest ever reported for educational interventions' (Black and Wiliam, 1998). In my own personal experience, I had been conscious of feeling somewhat lost and unfocused in situations where I was unable to get feedback about whether I was succeeding or not, and conversely much more in control in situations where I could get good feedback – even if it pointed out deficiencies. Again, I wondered whether I was unusual in this respect, or whether performance feedback was in general a necessary agent for improvement.

Early reading suggested that others had asked similar questions:

> … are teachers in effective schools more aware of what other teachers do in their classrooms? Do teachers in these schools have more opportunities to learn from other teachers (e.g. to observe, to engage in formal discussion) or to receive useful feedback from them? If teachers receive more feedback, what is the nature of the feedback? (Good and Brophy, 1986, p590)

> Providing clear and fair feedback to schools on their performance may be a feasible way to improve schools – letting schools improve themselves. (Fitz-Gibbon, 1992, p98)

However, there seemed to be no clear evidence derived from actually having tried it about the effects of giving such feedback. These speculations eventually coalesced into a hypothesis that could be tested: that giving teachers performance feedback might lead to improved performance. From this, a research design and methodology for testing the hypothesis soon followed, and the study was born.

The original aims for the research, as stated in my proposal for ESRC funding, were to answer four questions:

1. What kinds of feedback do teachers and schools use?
2. Are there any associations between particular uses of certain kinds of feedback and increased performance of students?
3. Is it possible to influence teachers' and schools' use of feedback?
4. If so, does such intervention result in any improvement in performance?

These questions remained close to the focus of the enquiry throughout.

About ALIS

The A Level Information System (ALIS) began as a small investigation into the relative performance of Mathematics and English departments in a handful of schools in the North East of England in 1983 (Fitz-Gibbon, 1985). Fifteen years later, ALIS and its sister projects run by the Curriculum Evaluation and Management (CEM) Centre at the University of Durham provide information about student achievement, attitudes and perceptions to over 5000 schools and colleges in the UK and beyond. A suite of projects can track students from Reception (age 4-5) to A level (typically age 18), and both the range of services offered and the number of institutions involved are still growing. Today there are two options for institutions wishing to join the performance indicator systems for advanced (i.e. A level, AS level and Advanced GNVQ) students: 'Basic ALIS' and 'Full ALIS'. The schools pay to join according to the numbers of students involved,[1] and then receive all the materials and analyses free.

In Basic ALIS, students complete a questionnaire at the beginning of their course (typically in year 12, i.e. aged 16). This asks them for information about their previous academic achievements (GCSE grades), their current programme of study, and various personal data such as sex, ethnic origin, home background, date of birth, etc. It also asks about their aspirations for future education and employment. In addition, institutions have the opportunity to use the International Test of Developed Ability (ITDA) which provides a measure of verbal and numerical ability for each student (see Fitz-Gibbon, 1996). At the end of the course, when the examination results become available, the institution receives a set of printouts showing overall characteristics of the students in each subject (average prior achievement, value added) and a list of individual students and their value added performance in each subject. 'Value added' is calculated as the residual in an Ordinary Least Squares regression model for each subject. The model uses linear regression of A level grade (coded as A=10, B=8, C=6, D=4, E=2, N=0, U=-2) on average GCSE grade (the average of all grades achieved, where A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1). In some subjects, where numbers are large, a separate regression equation is used for different syllabuses, or syllabus types. Put simply, therefore, the 'residual' or 'value

[1] In 1996 the costs for an institution with 100 students in the year group were approximately £450 for Basic ALIS and £900 for Full ALIS.

added score' is a measure of the progress made by an individual student, compared to that made by others in that subject in the same national cohort.

Institutions that opt for Full ALIS use the same questionnaire as those in Basic ALIS at the beginning of the course, but students complete a further questionnaire close to the end of their course. This asks about their personal circumstances (e.g. part-time work commitments), their reasons for choosing the course, their satisfaction with the support they have received, the conditions under which they have been studying and their experience of the course organisation. It also asks for their perceptions of the frequencies of a range of teaching and learning activities. Non-academic outcomes are measured, including their attitudes to the institution, to the subject, their participation in extra-curricular activities and the likelihood that they will continue in education. This information is fed back to the institution in the form of graphs of aggregated statistics and transcribed, anonymous, open comments.

Member institutions are able to request INSET from the CEM Centre and regular conferences are provided for users to exchange information and practice.

Context and relevance of the study

From the late 1980s, the number of schools involved in ALIS grew steadily, and the CEM Centre began to offer value added and attitude monitoring projects across the full range of schooling (Fitz-Gibbon, 1996). The growth of interest in value added may be attributed in part to educational reforms such as local management of schools and the 'incorporation' of FE and sixth form colleges, which gave institutions the power to make their own decisions to spend money on such monitoring projects. Possibly more important, though, was an increasing culture of accountability within education in the late 1980s and early 1990s and, in particular, the publication of school performance tables from 1992. Schools' examination results came to be popularly seen as a prominent indicator of their success. The publication of these 'League Tables' led, however, to widespread feelings of unfairness and increased demands for some kind of adjustment for context in the form of 'value added'. Despite having previously rejected the use of value added, the UK government commissioned the School Curriculum and Assessment Authority to produce a

consultation document on value added (SCAA, 1994), and subsequently a feasibility study (Fitz-Gibbon, 1997) was commissioned.

In embracing the notion of public accountability, however, policy makers gave little attention to the question of the extent to which schools were really responsible for their students' attainment. Moreover, the philosophy of confidential self-evaluation using performance indicators, which had characterised ALIS from the beginning, began to seem quite out of step with the politicians' agenda. Many of those who had been using value added measures of student performance for internal monitoring would no doubt have welcomed the move beyond 'raw' examination results, but at the same time have felt some anxiety about the uses to which such information might be put. Value added, which in projects such as ALIS had been seen by most users as a crude but nevertheless useful measure of student progress, seemed to have metamorphosed in the minds of politicians and public into an objective measure of teacher effectiveness.

The question of whether 'student progress' (or rather, statistical measures of 'value added') can really be equated with 'teacher effectiveness' is addressed in Chapter 2. In Chapter 3, the research evidence about the effects of giving feedback is reviewed, in the hope of providing a theoretical basis to justify giving 'performance feedback' in order to improve performance. The remaining chapters describe two experiments which sought to investigate the effects of giving particular kinds of feedback to teachers.

Even in simple and well understood systems it can be quite difficult to predict the effects of a particular action. Education, however, is far from being a simple system, and its history is littered with innovations and policies whose effects were not as intended. If we want to know what effect an intervention will have, we must try it; that is not simply the best way, it is the only way. The experiments described in this thesis are an attempt to do just that: to monitor the effects of giving teachers a particular kind of feedback. From the knowledge gained in this kind of study we can begin to have a basis for policy that is founded on evidence rather than speculation.

Chapter 2

School Effectiveness and School Improvement: The Relevance of the Available Research

This chapter presents a review of the research literature on school effectiveness and school improvement. It makes a number of criticisms of much of the available research, arguing that its contribution to knowledge about how to help schools improve is rather limited.

2.1 REASONS FOR A REVIEW

A review of the current state of the knowledge base in the research field(s)[2] of school effectiveness and school improvement seemed to be a necessary preliminary to embarking on the study that is described in this thesis. Two main reasons justify this.

In the first place, this study may be seen as located within those fields. In essence, it was an attempt at school improvement: an intervention in the work of a group of schools with the intention of improving their students' examination results.

[2] Traditionally, the twin disciplines of 'school effectiveness' and 'school improvement' research have proceeded rather more separately than together. However, recent attempts to integrate their

It also relied on much of the methodology of school effectiveness research: the use of statistical models of 'value added', including multilevel models, both in producing the feedback that was provided to teachers, and also in evaluating its consequences. It seemed important, therefore, to be aware of the existing knowledge within those fields in order to assess both the contribution they could make to the design and operationalisation of this study and the contribution any findings from the study might make to their collective knowledge base.

Secondly, it was felt that the question referred to in Chapter 1 (to what extent can statistical measures of value added be interpreted as measuring student progress, and how far is this, in turn, a measure of teacher, or school, effectiveness?) was an important logical precursor to this enquiry. Clearly this is a key question for a study that depends so heavily on the use and interpretation of value added feedback. It was important to know what the school effectiveness and improvement research effort could say about the interpretation of 'value added'. Moreover, the relevance of much of the research about feedback effects (which is reviewed in Chapter 3) seemed to depend on the assumption that what is being fed-back is in some way a measure of performance adequacy. Hence, the issue was not only whether the interpretation of 'value added' as 'teacher effectiveness' could be justified, but also whether those involved (i.e. the teachers in the study) would interpret it in that way.

A general review of the school effectiveness and improvement research literature was therefore conducted. However, it soon became apparent that the justification for equating 'value added' and 'teacher/school effectiveness' was in fact rather weak, despite the fact that the assumption of their equivalence seemed to be frequently – but not often explicitly – made. Moreover, the 'knowledge base' of school effectiveness research (SER) came to seem a rather shaky foundation on which to build attempts at school improvement. The 'general review' thus turned into a more critical examination of some of the issues on which the interpretation of 'value added' seemed to depend, and of the difficulties of applying school effectiveness findings to school improvement.

---

knowledge bases and methodologies (e.g. Gray *et al.*, 1996; Reynolds, *et al.*, 1996) provide grounds for optimism that they might one day be seen as a single research field.

2.2 LIMITATIONS OF SCHOOL EFFECTIVENESS 'KNOWLEDGE'

Eighteen individual issues are raised here, each rhetorically presented in the form of a question. These are grouped within six broad areas: (A) issues concerned with the choice of outcomes that are measured, (B) the validity of the statistical modelling used, (C) the identification of cause and effect, (D) the question of whether 'effectiveness' can be altered, (E) the lack of understanding of the mechanisms of 'effectiveness', and (F) the over-stated claims of the ubiquitous 'correlates of effectiveness'. All of these may be seen as criticisms of much of the available research in the field of school effectiveness and its limited applicability in achieving school improvement.

(A) Choice of outcomes

1. Are a range of cognitive and non-cognitive outcomes measured?

It has often been noted (e.g. Good and Brophy, 1986; Scheerens, 1992, p.69) that there is a tendency in school effectiveness research (SER) to emphasise pupils' cognitive outcomes as the most important – or indeed only – measure by which effectiveness is defined, and this may well be seen by many as giving appropriate weight to the most important aspect of schooling. However, these cognitive outcomes are often limited to a very narrow range of measures (e.g. performance in tests of native language and mathematics), and, moreover, are often restricted to the testing of low-order 'basic' skills (Cuban, 1984). It seems that the choice of outcomes is often driven more by convenience and availability than the desire to measure what is important.

2. Are the outcome measures used sensitive to teaching?

The tendency of many of the earlier SE studies to use curriculum-free standardised tests as outcomes has also been pointed out (e.g. by Madaus *et al.*, 1979), and it is now more common for researchers to recognise that if you are going to use a

test to infer the quality of teaching that has occurred, then it makes sense to test what has been taught. Nevertheless, there are still plenty of examples of research in which the sensitivity to teaching of the outcome measure may be thought questionable. Since SER has yet to demonstrate how much effect teaching can have on any outcome (see below), the best criterion we have for judging the appropriateness of an outcome measure is our impression of its face validity. These arguments about sensitivity to teaching – or to school influence – apply equally to non-cognitive outcomes such as attitudes, self-perceptions, social skills or behaviour.

3. Do the outcome measures match schools' objectives?

Even if it could be shown that the outcome measures used were within the control of schools (this issue is discussed below, p28), it would still be necessary to measure outcomes that reflect the educational objectives of the school. It would seem absurd to judge a school as 'effective' by measuring something that it had not tried to affect. A survey by Gray *et al.* (1986) found that no schools actually rated examination performance as unimportant, but there were substantial differences in the relative importance attributed to it. Of course, there is a political dimension here: if a particular outcome is decreed to be a measure of 'effectiveness' then it is likely that it will become a high priority objective. The extent to which schools should be free to set their own objectives is arguable, but any attempt to compare them on the basis of 'effectiveness' will require that they have common objectives, and, perhaps more controversially, that the same objectives apply to all students. Given the wide variety of schools that exist in any system, and the range of students' needs within many schools, it may be questioned whether this can be in the best interests of all students. Indeed, it could be argued that in England and Wales the recent increasing focus on examination performance, particularly in the higher grades used to calculate league tables, and recent dramatic increases in the number of pupils excluded from schools are, at least in part, cause and effect. It may well be that forcing all schools to adopt identical objectives for all pupils is not compatible with a comprehensive, inclusive educational system.

(B) Modelling

4. Is the 'residual' more than just the unexplained part of performance?

Measures of 'value added' are generally defined only by default. So called 'effectiveness' usually means that part of pupils' performance which cannot be accounted for by their intake characteristics: in other words, as a statistical residual. To assume that we can interpret such a variable – defined in terms of what it is not – would be unwise, even in a field where the theoretical relationship between variables were well understood. What is extraordinary is that this assumption is made at all, let alone that it is made almost universally and uncritically (with a few exceptions, e.g. Preece, 1989; Fitz-Gibbon, 1996), within a field where a sound understanding of how effective teaching and learning occur is almost non-existent.

It is hard to believe that no systematic attempt has ever been made to justify the validity of the statistical calculation of 'effectiveness', as defined in the 'residual' model. Its interpretability has rested entirely on common sense and plausibility – which would be fine as a starting point, but which are surely wholly inadequate as a basis for a mature research discipline, and even more so when used to make vital judgements about individual schools. It therefore seems important to know to what extent ratings of 'effectiveness', defined by such a 'residual' model, are validated by other independent and reliable measures of the effects of schooling, and to what extent expected relationships with other constructs are found.

5. How good are the 'control' variables?

The interpretation of 'value added' as 'effectiveness' depends heavily on the adequacy of the control variables used. Any relevant factors which are unmeasured, or measured unreliably, will make schools with better 'raw' performance seem more 'effective' than is fair. If no allowance at all is made for the intake characteristics of the students (i.e. if 'raw' outcomes are used), it would be widely felt that this would not measure the effectiveness of the school: the best school with a disadvantaged intake could never perform as well as a mediocre school with the head start of a more able population. On the other hand, if theoretically perfect control variables – which

measured with complete accuracy every relevant aspect of the individual students and all the contextual factors which were outside the control of the school – were used, it might be thought that the residual did indeed measure the effect of the school. The reality, of course, will be somewhere between the two extremes, but studies which, for example, have no measure of prior achievement but adjust only for socio-economic status (SES) may be closer to the former than the latter (see Gray *et al.*, 1990). In practice, therefore, any measure of value added which we calculate may be thought of as a an attempt to measure 'pure' value added which is biased towards unadjusted (raw) performance.

We are still a long way from being able to say that we know what a complete set of control variables would look like. Typical value added models of school effectiveness, with the best data, are able to account for only about half the variance in individual pupils' performance (Tymms, 1996; Gray, 1995). Tymms (1996) has argued that the complexity of schooling is such that the remaining half may be in principle unpredictable. However, it seems at least plausible that individual characteristics such as motivational style or self-esteem, if measured appropriately, might account for a further part. The need for further research to explore the relationships between motivational style and performance has been highlighted by Galloway and Rogers (1994). Certainly, motivation and self-esteem have been shown to be associated with learning gains, independently of past achievement (Marsh, 1990; Zimmerman *et al.*, 1992; Fortier *et al.*, 1995) and are elements of most theories of learning, but form no part of the standard repertoire of control variables in SER. Clearly there are difficulties with obtaining valid and reliable measures of these characteristics, and it may be that even if such were available, the amount of additional variation accounted for might not be large. What is certain, though, is that any variables that are used will be measured with less than perfect reliability, and that unreliability in the control variables (or the omission of relevant variables) will result in residuals being biased towards the raw scores. In view of this, it is perhaps surprising that the reliabilities of the variables used are seldom reported in SER.

6. Do the variables used really explain variations in performance?

The broader issue here concerns the criteria that are used to decide on the inclusion of a variable in the statistical model. Typical SER seems happy to include a variable if it is easily measured (or, better still, already available) and accounts for a statistically significant proportion of the variation in performance. It is rare, however, for any consideration to be given to the theoretical significance of that variable. The issues of why and how it might be related to the outcome in question are unaddressed. These issues are important, though, if are to try to understand the reasons why some schools appear more 'effective' than others, rather than simply reporting the fact.

An example of such ungrounded modelling is found in the use of variables such as 'sex' or 'ethnic origin' which 'explain' (in the statistical sense) part of the variation in outcomes, but which do not explain differential performance in any true sense – unless it is argued that it results from purely biological differences, or from unfair discrimination. These variables are therefore being used as a proxy for some unmeasured characteristic with which they are associated, and which would genuinely explain why some individuals perform better than others. Presumably if this characteristic were identified and adequately measured it would account for significantly more of the outcome variance than the crude proxy. It could therefore be argued that the inclusion of such variables as controls is an indication that better controls could improve the model. This clearly points to the weakness of using purely statistical, rather than logical, criteria for including a variable as a control.

A further example of an inadequately conceptualised variable is the ubiquitous SES (socio-economic status). The fact that this variable is operationalised in different ways (for example as parents' occupations or family income – 'free school meals') in different studies is in itself interesting: can we infer different perceptions of the mechanisms by which home background affects school learning, or is it simply a question of what data were easily available to the study? Perhaps if we had a better understanding of which home background factors were important in influencing achievement, we would not only be able to formulate better value added models, but, more importantly, we might be able to do more to redress the inequality of disadvantage. SER could have a significant part to play in providing this understanding if its thinking were clearer.

7. Are genuine explanatory variables 'controlled' away?

One final way in which the validity of the value added definition of effectiveness could be compromised would be if true effectiveness were actually related to one or more of the intake variables whose effects are statistically allowed for in the model. For example, it is quite possible that teachers who work in schools in relatively disadvantaged areas may differ in significant ways from those in areas of greater advantage. Let us suppose, for the sake of argument, that those who take on the challenge of working in an environment of poverty and social dysfunction are generally more effective (in terms of the results they would achieve with equally matched groups of pupils in comparable circumstances) than those in more affluent schools. Because the 'residual' model compensates for the effect of pupils' socio-economic status, it will also unwittingly wipe out at least part of a genuine difference in effectiveness. Within a multilevel model this 'overcompensation' would contribute to a compositional effect (i.e. the apparent effect of a school's average SES, over and above the effects of SES on performance at the individual pupil level). This issue is clearly a complex one, but once again points to the need to take account of the expected theoretical relationships among variables, and not just their statistical relationships.

(C) Causality

8. Do schools really make a difference?

Following the findings in early studies that schools appeared to be far less significant than socio-economic factors in accounting for differences in student performance (Coleman, *et al.*, 1966; Plowden Report, 1967), it became fashionable to present later reports of school effectiveness research with titles such as 'Schools Make a Difference' (Teddlie and Stringfield, 1993; Brookover *et al*., 1979; Mortimore *et al.*, 1988). It is interesting to note, as an aside, that the percentage of variance attributable to the school in these later studies is comparable to that found in the earlier studies (Bosker and Scheerens, 1989); only the interpretation has changed. However, claiming that 'Schools Make a Difference' is quite different from claiming simply that

schools are different - a much less spectacular claim, but much more in keeping with the evidence available. The question of causality is crucial.

It could theoretically be argued that differences in pupil outcomes are purely a result of (possibly unmeasured) differences at intake, or even of unpredictable (chaotic) interactions among well measured factors. Indeed, it has been claimed that apparent differences are in fact 'statistical mirages' (Preece, 1989, p.65), and it has even been suggested that school outcomes may be in principle unpredictable, within certain limits (Tymms, 1996). School effectiveness research must therefore demonstrate causality: that apparent differences between the performance of students in two schools are genuinely a result of attending those schools and not simply due to unmeasured initial differences, or chance variation. Strictly speaking, only if randomly allocated groups of children were consistently found to perform differently in different schools could we be sure that the difference was *caused* by attending that school. Of course, there are real difficulties with conducting true experiments, and it may be argued that the issue of practical significance is really control rather than causality (see point 9, below). Moreover, sufficiently secure causal attributions can, under certain conditions, be made without experimental data (Holland, 1986). The question of causality is important because our conceptualisation of 'school effectiveness' depends on an understanding of this issue.

Some prominent school effectiveness researchers have acknowledged the absence of evidence about causality (e.g. Fitz-Gibbon *et al.*, 1989, p.144; Scheerens, 1992, p.71; Gray *et al.*, 1995, p.221; Reynolds and Stoll, 1996, p.104), but the impression often gained – even where the issue is mentioned – is that it is something of a technicality, rather than a fundamental flaw in the methodology of school effectiveness research. Scheerens (1992, p67) advocates 'broadening the arsenal of research methods' to go beyond the typical correlational study and use quasi-experimental and even truly experimental designs, alongside naturalistic case studies and alternative approaches such as computer simulations. Only by extending the repertoire can we develop a sound understanding of school effectiveness and thereby throw light on the extent to which apparent school 'effects' are indeed causal.

(D) Control

9. Can schools alter their 'effectiveness'?

A number of commentators have questioned whether schools are really responsible for – or able to influence – their 'effectiveness'. This point is, of course, closely related to the previous one. In the words of Cuban (1984), 'no one knows how to grow effective schools', and, more recently, Tymms (1996), 'The answer to the essential question "How can we improve our schools?" is no clearer now than it was a decade ago'. The same sentiment is put in somewhat more understated terms by Gray *et al.* (1996, p177): 'there are sizeable gaps in our understanding of how to turn knowledge about school effectiveness into enhanced strategies for school improvement.'

This issue is evidently crucial if the intention of the school effectiveness research effort is ultimately to improve schools, rather than simply to measure them. It also has fundamental implications for the judging of individual schools based on their effectiveness, or for requiring them to set targets (DfEE, 1997). It makes no sense – quite apart from being grossly unfair – to praise or condemn a school for its apparent effectiveness if there is no good reason to believe that anything that school could have done would have made any difference to it.

10. Has school improvement research shown that schools can improve?

Given the need for evidence about the extent to which schools can actually make changes to bring about improvement, and the abundance of school improvement initiatives throughout the world, it might seem that evaluation studies of these initiatives could be a rich source of such evidence. Unfortunately, with a few well designed exceptions (e.g. Reynolds *et al.*, 1989), these evaluations often prove inadequate for this purpose, for one or more of a number of reasons (Scheerens, 1992, p.56; Reynolds and Stoll, 1996, p105). In the first place, there is no general agreement about what actually constitutes 'improvement'. As Gray *et al.* (1996, p178) have said, 'It will continue to be difficult to make worthwhile assessment of the results of school improvement efforts for as long as researchers and practitioners

remain reluctant to assess the impact of their activities on pupils'. The evidence about what constitutes improvement is often more concerned with changes in the perceptions of those involved, rather than observed behaviour or measured performance of either staff or students. This is not to say that perceptions are not important, but they are notoriously poorly associated with the more 'objective' measures which are generally used in SER. It may be quite hard to know how to interpret the perceptions of people in schools who may well have been instrumental in initiating the improvement and have probably invested considerable commitment in making it work.

Secondly, evaluations are often poorly controlled; in other words, they fail to rule out alternative explanations for the 'improvement' seen. For example, improvement initiatives are often launched within changing educational systems, and without an adequate control group it is impossible to judge what might have happened without the initiative. Schools who become involved in such initiatives are inevitably volunteers, and it is certainly arguable that, by the time the school's management have identified the need for improvement and made the commitment required by a particular improvement programme, the actual details of that programme are pretty much irrelevant: from that point they would probably have improved whatever was done. Another assumption widely made in poorly controlled school improvement evaluations is that intakes have remained constant. It may in fact be that the easiest way to improve a school's performance is to improve its intake, but for most people this could not reasonably be described as 'school improvement'. One further spurious way to apparently bring about improvement would be to start with a school whose performance was poor in a given year and rely on the natural year-to-year variability to deliver better performance as a result of regression to the mean. Without adequate controls, it would certainly seem that 'failing' schools were easier to 'improve' than others.

Thirdly, evaluations of school improvement initiatives have tended to rely on short term outcomes and are therefore unable to provide any evidence about the sustained effects on performance. Given the variability of even the best available measures of school effectiveness from year to year (Sammons *et al.*, 1996), as well as the cost of implementing improvement initiatives, it is important to know the longer term effects.

Many of these weaknesses are being addressed in current studies which draw on the previously separate fields of school effectiveness and school improvement (Reynolds and Stoll, 1996; Gray *et al.*, 1996) and the emergence of this 'merged paradigm' is to be welcomed. It is hoped that future critical evaluations of well controlled interventions will enable us to say more than simply that a school has improved, but to achieve a more sophisticated understanding of how, why and how much.

(E) Understanding of mechanisms

11. Why are some schools more 'effective' than others?

This question has been raised by, for example, Willms (1992, p64): 'after two decades of serious effort, [researchers] have made little progress in determining *why* some schools are more effective than others'. School effectiveness research has been characterised as being like a 'fishing expedition for significant correlations' (Scheerens, 1992, p.67), which gives few answers to the question of why they exist.

12. Are 'effects' sought where they might be expected?

A number of examples have already been given of ways in which SER has proceeded opportunistically and somewhat blindly, rather than being guided by a clear theoretical rationale. Lack of theory may also have had a part to play in the continued searching for school level factors associated with effectiveness, despite the fact that learning takes place primarily in classrooms, and may therefore be expected to be influenced more by classroom level factors. Evidence that different departments within a school may have quite different 'effects' has been available for some time (Fitz-Gibbon *et al.*, 1989; Tymms and Fitz-Gibbon, 1990) and has recently been more widely acknowledged (e.g. Sammons *et al.*, 1996). An observer could be forgiven for gaining the impression that the assumption that schools had homogeneous 'effects' was convenient when data on different kinds of outcomes were not available, but was found to be untrue when they were. It might be wondered whether if data about the performance of students taught by different teachers become available, the construct

of 'subject effectiveness' might not also dissolve. Of course, the question of homogeneity of effects is an empirical one, but researchers will not ask the question (i.e. go to the trouble of collecting the required information) unless they have some *a priori* reason for believing that a particular phenomenon might be found. Such a reason can only emerge from a well grounded model of the mechanisms by which different factors interact, and the absence of such a model to guide the research enquiry makes progress far less likely. The design of the COMBSE project (Fitz-Gibbon, 1985) and its successor, ALIS (Fitz-Gibbon, 1992) to provide separate 'value added' analysis for each subject provides a good example of research guided by such a combination of closeness to the data and attention to the relevant mechanisms.

Some writers, such as Scheerens (1992) and Creemers (1994) have attempted to produce models of educational effectiveness. Creemers (1996) reviews such attempts, all of which focus on the learning of individual students and, in particular, Carroll's (1963) model of learning. Creemers emphasises the need to place student learning within a multi-level structure, identifying factors at student, classroom, school and context level that may influence what is learnt. For each factor, the empirical evidence in its support is summarised.

However, some of the limitations of studies that have sought these factors are not addressed, and these will be considered now.

(F) Correlates, not causes

13. Is there genuine consensus about the correlates of effectiveness?

It is now something of a tradition in reviews of school effectiveness research to begin by listing sets of characteristics of schools which have been repeatedly found to be associated with 'effectiveness'. Commonly cited are Edmonds's (1979) 'five-factor model', Purkey and Smith's (1983) model with eight factors, Mortimore *et al.* (1988), who expand the list to twelve, and Sammons *et al.* (1994) who reverse the trend by bringing the number of factors back down to eleven. It will be argued here that the obsession of school effectiveness research with reproducing these 'effective school correlates' (Levine and Lezotte, 1990) is an obstacle to real progress.

It is often asserted that there is broad consensus about the existence of these 'effective school correlates', despite a number of inconsistencies (Creemers, 1996). However, although many of their formulations may be broadly similar, the precise operationalisation of each factor is often peculiar to the particular study in which it is found. It would be hard for proponents of these 'n-factor models' to refute the argument that when they repeatedly find, for example, correlations between 'strong educational leadership' and effectiveness that they are not capitalising, at least to some extent, on chance associations and ambiguities in the definition of the factor. A typical correlational school effectiveness study will measure a range of process factors and report any statistically significant associations with effectiveness. An uncritical reviewer of such studies will systematically seek overlap between the meanings of these factors. Thus the less well defined a factor is, the more likely it is to be counted as a confirming instance of a general result. There may therefore be a significant bias in such reviews towards finding a consensus. While it is unlikely that the consensus is entirely spurious, it is also unlikely to be as strong as it appears.

14. How well have the results of different studies been integrated?

There are, however, even stronger arguments against the synthesis of the research evidence by means of such 'vote counting'. It has long been shown that simply counting the number of studies which find a particular phenomenon and balancing them against those that do not can give a result opposite to that supported by the data considered as a whole (Hedges and Olkin, 1980), even supposing that the 'file drawer problem' (Rosenthal, 1979) has not rigged the 'vote' by making studies which failed to find the effect relatively harder to find. Moreover, it has even longer been argued (Tukey, 1969) that simply to report the existence of an association, with no measure of its size is to discard the main import of the data:

> If, for example, elasticity had been confined to "When you pull on it, it gets longer!", Hooke's law, the elastic limit, plasticity, and many other important topics could not have appeared. … Measuring the right things on a communicable scale lets us stockpile information about amounts. (p.86, 80)

Tukey (*ibid.*, p.89) goes on to argue that, because of their sensitivity to the amount of variability in each measure, correlation coefficients are less appropriate than regression coefficients for quantifying the size of causal effects.

Unfortunately, the information typically presented in reports of school effectiveness research does not often allow this quantification to take place. Scheerens (1992, p.55) describes the attempt to review and synthesise twelve key studies in school effectiveness research as a 'sobering experience'. He stresses the need to examine the original studies, rather than the numerous reviews (of a relatively small number of studies) which have repeatedly summed up the same correlates of effectiveness in a rather uncritical and superficial way. Scheerens observes that as, 'basic quantitative data are often missing from the publications … the attempt to reach a quantitative synthesis was abandoned' (p55). This is a particularly devastating indictment of SER, since only by conducting this kind of synthesis (i.e. by meta-analysis) can the evidence from different studies be satisfactorily combined.

15. Could the relationships be more complex than those commonly sought?

The prominence of these 'effective school correlates' in the literature may also tend to constrain the search for relationships between school features and effectiveness to a search for linear relationships. It seems *a priori* far more likely that some relationships will be 'threshold effects' (Tymms, 1996), in other words that effectiveness may be reduced if some characteristic fails to reach a particular value, but may not increase appreciably beyond that point. Equally, one might expect that many relationships would have an inverted U shape if plotted over sufficient range (Fitz-Gibbon, 1985, p.51; Preece, 1989). For example, effectiveness may well increase with 'frequent monitoring of student progress' (Edmonds, 1979), but perhaps only up to a certain point; beyond that, effectiveness might be expected to decline. It would certainly be of value in such a case to know the amount of the characteristic that gave optimal effectiveness. It is also not necessarily the case that all relationships between school characteristics and effectiveness will apply equally to all groups of students. Preece (1989, p.62) cites an analysis by Chapman in which the correlation between examination performance and teachers' frequent use of dictated notes was positive for lower achievers, and negative for higher achievers.

16. Has a focus on the outcomes being measured been mistaken for a new factor?

The issue of the lack of theory in school effectiveness research has been mentioned already, and, in particular, the need to recognise the multi-level structure of schooling in searching for relevant factors. An example of how this lack of theory may lead to a misunderstanding of the relationship between a school characteristic and effectiveness has been given by Scheerens and Creemers (1989). They question whether the finding that an 'emphasis on basic skills' is correlated with effectiveness may owe more to the fact that basic skills are frequently taken as the measure of 'effectiveness' than to the genuine importance of this factor. They suggest that the factor more likely to be generally associated with effectiveness is the extent to which goals are congruent with measured outcomes (the absence of 'goal-measurement disparity', p267). A similar point has been made by Gray *et al.* (1986, p92) who question whether 'some part of the apparent differences in results that emerge between schools arise not so much from differences in effectiveness as differences in objectives'.

It might seem rather disappointing if one of the main results of SER were to be recognised as the finding that schools often achieve only what they set out to achieve: schools whose focus is on the outcomes used to define 'effectiveness' are more 'effective' than those who are less focused on these outcomes. However, this may actually be quite an important finding. Indeed, it may be that if any of a number of recent UK government initiatives such as target-setting – or even the publication of school performance tables – do actually result in genuine improvements it will be largely owing to their effects on the focus of people's activity. Chapter 3 provides some discussion of this issue.

17. Has correlation been confused with causation (again)?

Finally, we must return to the question of causality. It seems obvious enough that 'high expectations for students' (Edmonds, 1979) is as likely to be a result of high standards as its cause, and, in relation to this particular factor at least, this point has often been made (e.g. Scheerens, 1992, p.80; Reynolds and Stoll, 1996, p.104). Nevertheless, the same argument applies equally to all the other correlates. Indeed, it is quite possible to produce explanations – albeit, sometimes quite tortured – of every

single correlation in which either the causality is reversed (i.e. the existence of high effectiveness causes the characteristic to be found) or both phenomena are caused by a third factor. It has also been argued (e.g. Scheerens and Creemers, 1989) that the factors themselves are causally interrelated. The fact that everyone knows 'correlation is not causation' does not appear to have prevented the existence of acres of print which assumes (implicitly or explicitly) that schools which seek to take on the characteristics associated with effective schools will thereby become effective. Convincing evidence that this is so, however, is almost non-existent.

18. Does the research evidence take us beyond common sense?

Anyone who looks at a list of 'correlates of effectiveness' – whether in five, eight, twelve or some-other-number-of-factors form – will surely be struck by how obvious they all seem. The knowledge that, for example, 'unity of purpose' and 'an orderly atmosphere' are more likely to be associated with effectiveness than disunity or disorder falls some way short of justifying the huge endeavour that is school effectiveness research. Moreover, since the factors are generally presented in the form of a long list with no obvious order of importance, one could be forgiven for questioning their usefulness as a guide to action. Any school that is trying to do its best in an intelligent way will more or less be doing all of them already – to a greater or lesser degree.

Of course, the fact that a research field produces results in broad agreement with common sense is in itself not an argument against it – quite the reverse. However, it may be that the real value of its contribution to knowledge lies in those areas where it appears to conflict with previously held intuition. One way in which SER might have contributed more in this respect is by ruling out other equally obvious factors that were found not to be associated with effectiveness. This point was well made by Rutter *et al.* (1979) in defending their research against the possible accusation that it said only what was obvious. Although subsequent studies have often – but by no means always – listed all the factors tested, whether or not associations were found, the importance of excluding irrelevant factors from consideration has generally been overlooked, particularly in the reviews of SER. Once again, the only sound way to

integrate the evidence about the importance of all the factors tested would be in a meta-analytic quantitative synthesis.

2.3 CONCLUSIONS FROM THE REVIEW

A number of issues have been raised which suggest that the interpretation of 'value added' is extremely problematic. In particular, how much of the responsibility for students' value added performance should be attributed to teachers is by no means clear. More pragmatically, it is far from obvious that the entire school effectiveness research effort can really justify offering any advice to teachers and schools about how they might improve that performance.

Chapter 3

Can Feedback Improve Teaching? A Review of the Evidence

This chapter returns to the general question of the effects of giving people feedback. It presents a review of the social science literature with a view to identifying the conditions under which giving feedback to teachers may be expected to result in improved performance.

3.1 INTRODUCTION

Increasingly prominent practices in education such as inspection, the use of quality assurance procedures, publication of a variety of performance indicators and appraisal are all in part motivated by the belief that feedback is somehow good for us. Indeed, there can be few statements in social science more likely to gain agreement than the notion that giving feedback can improve a person's performance on a task, and few which have been the subject of more research. However, a closer examination of the evidence reveals a far more complicated picture: feedback is by no means always beneficial in its effect, and identifying the conditions under which it may be expected to improve performance is far from straightforward.

In a review of what was already then over fifty years of published research on the effects of feedback on performance, Ammons (1956) concluded:

> Almost universally, where knowledge of their performance is given to one group and knowledge is effectively withheld or reduced in the case of another group, the former group learns more rapidly, and reaches a higher level of proficiency. (p283)

Some thirty years later, Pritchard, *et al.* (1988) could state:

> ...the positive effect of F[eedback] I[ntervention] on performance has become one of the most accepted principles in psychology. (p338)

Despite the obvious plausibility of this principle, however, it does not appear to be borne out, at least in so simple a form, by the evidence from experiments. In a meta-analysis of 131 studies (607 effects) on the effects of Feedback Interventions (defined by them as 'action(s) taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance'), Kluger and DeNisi (1996) found that although the average effect was moderately positive (weighted mean effect size[3] 0.41), over 38% of the effects were negative and the mode of the distribution of effect sizes was zero. They conclude:

> FIs do not always increase performance and under certain conditions are detrimental to performance. (p275)

Similar results have been found in other reviews and meta-analyses (e.g. in Bangert-Drowns *et al.*, 1991; Locke and Latham, 1990).

It seems important, then, to examine the evidence with a view to identifying the conditions under which giving feedback does result in improved performance. If a general theory can be found which enables us to generalise those conditions and to understand the mechanisms involved, then so much the better.

---

[3] Effect size is a measure of the difference between the performances of experimental and control groups, expressed as a proportion of standard deviation. Where an average effect size is calculated from a number of studies (e.g. in meta-analysis), individual results should be weighted so that large studies contribute more to the overall average (Glass, McGaw and Smith, 1981).

3.2 ANALYSIS OF THE RESEARCH EVIDENCE

Limitations of the existing research

The role of theory

Despite the existence of a large quantity and range of literature on feedback and performance, systematic attempts to identify which variables may be significant in mediating the effects of feedback on performance are most notable by their absence. Existing research often seems more concerned with establishing or rejecting a particular theory than with seeking the conditions under which given phenomena occur (Greenwald *et al.*, 1986).

Kluger and DeNisi (1996) claim that there is no single universal theory of the mechanism by which feedback affects performance; indeed, they cite this as one of the reasons why the perception that feedback always has a positive effect has been maintained, despite the mixed empirical evidence. There are, however, a number of psychological theories that relate to some aspect of the interaction between feedback and performance, generally dealing separately with either motivation or learning. Research on motivation is frequently constrained by its particular theoretical orientation (Bong, 1996) and often makes no mention of any resulting effects on performance. It is also often assumed that motivation can be treated as a global characteristic, despite evidence that motivational style is more a product of situational than individual variables (Galloway *et al.*, 1996). Research on learning is often concerned with very specific and low level learning which takes place in the space of a few hours in a laboratory, and is therefore of doubtful relevance to performance in a complex activity such as teaching.

There are some more general theories, including Kluger and DeNisi's own (1996) *Feedback Intervention Theory* and Locke and Latham's (1984, 1990) *Theory of goal setting*, about which more will be said below. However, the impression gained from reading the literature is that the role of theory seems to be more to provide *post hoc* explanations of a complex tapestry of apparently anomalous results, rather than to enable clear *a priori* predictions to be made.

Semantic drift: comparing like with like

A further difficulty arises from the wide range of interpretations given to the words 'performance' and 'feedback', together with a variety of other experimental variables, some of which are acknowledged, some are not. It could certainly be argued, for example, that the mean effect size, quoted above from Kluger and De Nisi (1996), is the meaningless result of a comparison of like with unlike. Further examples of this difficulty may be found in studies such as Cohen's (1980) meta-analysis of the effects of feedback, which concludes that

> ... student-rating feedback has made a modest but significant contribution to the improvement of college teaching. (p336)

However, the outcome measure used here is the change in the instructors' behaviour as rated by the students, and therefore may be unrelated to change in student performance, or any other measure of teaching performance. Even 'student progress', on which Cohen estimates the effect of feedback as represented by an effect size of 0.30, is defined in his meta-analysis in terms of students' ratings of their own progress. This, again, may be quite different from progress as measured by achievement tests. Similarly, Brinko (1990, 1993) appears to equate feedback effectiveness with a number of outcomes, including teachers' behavioural or attitude change, but the definition is unfortunately not made clear.

Ecological validity: transfer of results across contexts

If our concern is to make predictions about the effects of particular kinds of feedback on a specific group of teachers in particular institutions, then we are almost entirely dependent on generalising results from other settings. Hardly any work has been done specifically on the effects of feedback on teachers. How far results can be transferred from one setting to another is – in the absence of any empirical evidence – largely a matter of judgement. However, a small number of studies have been found that have specifically investigated the effects of providing performance feedback to schools. These are briefly summarised now.

Studies that have specifically investigated giving performance feedback to teachers

Cohen (1980)

Cohen's meta-analysis has already been mentioned in relation to the importance of being clear which outcomes are measured. In addition to the effects cited above, however, some of the studies included in the meta-analysis recorded the effects of the feedback on student attitudes, and some recorded the effects on student achievement. These are therefore of more relevance to the present investigation.

All the studies analysed by Cohen concerned the effects of giving student-rating feedback to college teachers in the US. The same three studies recorded the effects of the feedback on students' attitudes towards the subject and on student achievement. All three provided the feedback in the form of a 'consultation' and one also provided another treatment group with just printed feedback. Thus a total of four effects were calculated for each outcome. Two of the three studies (and three of the four effects) allocated teachers randomly to treatments, the other used covariance analysis.

In terms of the effect of the feedback on student attitudes, all four comparisons favoured the feedback group and the overall effect size was 0.42. For student achievement, three of the four favoured the feedback group, while one showed better performance by the students whose teachers had not had the feedback. The overall effect size on student achievement was 0.19.

Brinko (1990, 1993)

Brinko conducted a review of the literature on feedback effects, with a view to applying its findings specifically to the effects of giving feedback to teachers. Her review ought therefore to be an extremely valuable precursor to this study. Unfortunately, however, the outcomes used to define the 'effects' of feedback are not made clear, so it is quite hard to divine exactly what is being claimed, but seem generally to be concerned with the teachers' attitudes or behaviour. No numerical estimates of effect sizes are given.

Brandsma and Edelenbos (1992, 1998)

Brandsma and Edelenbos conducted an experiment in The Netherlands in which specific forms of training were given to school principals and mathematics teachers, and the effects on their students' performance evaluated. Part of the training the principals received was in the interpretation of value added performance data for their own pupils. They were also trained to implement other practices identified as optimal from the school effectiveness research literature. The teachers were trained, at somewhat greater length than the principals, to structure their teaching and to provide feedback to their pupils. Performance feedback was therefore only a small part of the experimental intervention.

The results were somewhat disappointing in that neither the principals' or the teachers' training had any appreciable effect on subsequent student performance, including on a retention test a year later. Moreover, the effect of both interventions together appeared to be, if anything, slightly negative. Whether this can be taken as evidence about the effects of feedback is questionable, since feedback to individuals on their own performance was only a small part of the interventions. However, it is interesting that the best advice from school effectiveness research appeared to lead to no benefit at all when it was applied in a well-evaluated school improvement initiative.

Tymms (1995, 1997a, 1997b)

Tymms (1990, 1995) has argued for the view that giving performance feedback to schools can improve performance and has conducted a number of experiments to investigate the effects of such feedback. The first of these (Tymms, 1995) was concerned with teachers' responses to different forms of the feedback sent by ALIS (a long or a short version) and to receiving an invitation to attend an in-service workshop. Overall, the two kinds of feedback did not lead to significant differences in teachers' attitudes or self-reported behaviour. There were some differences in the responses for teachers of different subjects, but given the low – and differential – response rate, these are hard to interpret unequivocally. Sending an invitation to the in-service workshop did appear to lead to more positive attitudes towards ALIS.

Tymms' second experiment was conducted as part of the Value Added National Project (Fitz-Gibbon, 1997). Tymms (1997a) randomly allocated 257 primary schools to receive feedback about their value added performance in a number of different ways. Once again, those who were invited to INSET on the feedback were more positive about it. The form of feedback also had an effect on pupils' subsequent Key Stage 2 results, with those who received the data in the form of tables subsequently achieving slightly better results than those who were sent graphs. The difference was small (0.073 in terms of average KS2 level, adjusted for previous KS1 average and school percentage free school meals), but corresponded to an effect size of 0.2. Interestingly, most of the difference was accounted for by improvement in the level achieved in English.

Tymms' final experiment (1997b) was conducted with schools in the Performance Indicators in Primary Schools (PIPS) project, the primary phase in the suite of indicator systems provided by the CEM Centre at the University of Durham. It compared the 1997 performance of schools who had joined the PIPS project at its beginning in 1993 with those who joined as part of a whole LEA in 1996. The differences between the two groups were measured by their adjusted performance in Key Stage tests and pupils' attitudes. Both had effect sizes of about 0.1, in favour of those who had been in PIPS from the beginning. Although the initial invitation to join was sent to a random sample of schools in the LEA and the majority (nearly three-quarters) of those invited did join, when calculated from the difference between those who were invited and those who were not, the effect size shrank to zero.

Framework for analysis of other studies

The vast bulk of the available evidence on feedback effects comes from contexts other than those in which performance feedback is given to teachers. In order to judge the extent to which findings may be transferred across contexts, it is important to try to understand which contextual variables mediate the effects of feedback, and in what ways.

The following list of conditions on which the impact of feedback on performance may depend is drawn from a wide range of sources. They are grouped for convenience into three types. The first consists of factors that are primarily

characteristics of the type of task on which performance is being measured, and the context in which the task is performed. These factors are thus determined by the type of 'performance' that it is desired to influence and are not manipulable once that choice has been made. They are, however, of interest, since they may help an understanding of why feedback has or does not have a particular effect, and enable a better prediction of effects to be made. Secondly, there are *characteristics of the particular feedback* that is given, or of the way it is given. These factors can generally be manipulated by the feedback provider and it is therefore important to know which particular manipulation of them may be expected to have maximum impact on performance. Third and finally, are individual *characteristics of the person receiving the feedback*. This category includes any factors which vary at the level of the individual. Some of these may well be manipulable, although others may not. Nevertheless, it is important to know which feedback effects are likely to be generalisable to all recipients, and which may improve the performance of some more than others.

Evidently, these distinctions are not clear-cut. The way a person perceives certain feedback could depend on the individual as well as on the way it is presented, and there seem to be many interactions between factors. Nevertheless, they provide a convenient structure for analysis.

Significant variables 1: Characteristics of the task

Complexity

Much of the research on the effects of feedback relates to tasks performed as part of a laboratory experiment. These typically provide tasks of short duration which are relatively simple in structure, although the tasks may be difficult. In a review of research on informational feedback, Mory (1992) observes that:

> ...tasks involving higher cognitive processes, such as rules and concepts, do not produce the same feedback results as do rote memorization tasks such as verbal information. (p13)

An analysis of the relationship between the complexity of a task and the effect of goal-setting on performance (goal-setting effects are inextricably linked with feedback effects – see 'Relationship to goals', p46, below) is presented in Locke and Latham (1990), based on a meta-analysis by Wood, Mento and Locke (1987). Their sample shows 'a strong bias towards more simple tasks such as brainstorming, perceptual speed and toy-assembly tasks' (p218). The mean effect of goal-setting on performance (effect size corrected for reported reliability) for the least complex tasks was close to 0.8. However, it fell rapidly, and for the more complex tasks was just above 0.4. Even so, this is still a substantial effect. Locke and Latham (1990, p317) argue that the effectiveness of feedback in complex tasks depends on its effects in strategy development. Simple 'outcome feedback' may be ineffective unless it is supplemented by specific and diagnostic feedback as well as guidance on choice of strategy.

Balance between demands of effort and ability

Many of the tasks used in experiments are such that increased effort may well lead directly to improved performance on the task. In such a case the role of feedback may simply be to enhance motivation and thus increase performance. However, in many real-world tasks, such as teaching, it is not clear that simply trying harder will improve performance. Indeed, it is true in general that the relationship between motivation and performance is modelled by an inverted U: there is an optimal level of motivation to produce maximal performance, and increasing it beyond this level is likely to reduce performance (Costanzo *et al.*, 1992).

Some feedback that has been reported to improve performance is essentially 'compliance feedback' rather than 'performance feedback'. An example of this is Archer-Kath *et al.*'s (1994) study in which feedback given to children on the use of particular social skills in group-work resulted in increased use of those behaviours and, presumably as a consequence, improvements in academic performance and attitude.

A distinction may need to be made between tasks in which increased motivation and effort, arising from feedback, might be expected to improve performance, and

those in which motivation and effort could increase without any such effect. However, it is not clear into which category teaching would fall.

Availability of other information or instruction

It seems reasonable to expect that when a particular type of performance feedback provides the only information a person has with which to improve their performance, then the improvement which is attributable specifically to that feedback is likely to be greater than if a large amount of good quality information is available. Equally, in a situation where feedback is accompanied by other forms of guidance or instruction it seems likely that the measurable effect of the feedback will be less than if these alternative aids to learning are not available. The implications of this are perhaps of more relevance to those who are designing or interpreting experiments to show feedback effects than to those whose aim is to improve performance, since isolating a particular piece of feedback may make it easier to identify it as the cause of improvement, but is unlikely to maximise performance.

This plausible expectation is supported by Bangert-Drowns *et al.* (1991) who found that the use of a pre-test appears to reduce the effect size of a feedback intervention, and concluded that,

> feedback is more important when the content is more complex and when the student is given fewer cues, organizers and other instructional supports. (p233)

In the context of the classroom, teachers generally have a good deal of information about the quality of their performance from a wide range of sources as well as a variety of forms of support and training. It may therefore be expected that the impact of providing additional performance feedback will be small.

Significant variables 2: Characteristics of the feedback or the way it is presented

Relationship to goals

An important theory in the field of organisational behaviour is that of *Goal Setting* (Locke and Latham, 1984, 1990). According to this theory, providing

feedback *per se* does not improve motivation or performance, but it will do so if it leads to higher goals being set, or greater commitment to existing goals. Locke and Latham (1984, p15) quote a number of experimental interventions which have:

> ...demonstrated that goal setting in industry worked just as well as it did in the laboratory. Specific, challenging goals led to better performance than easy or vague goals such as 'do your best' , and feedback motivated higher performance only when it led to the setting of higher goals.

It does not seem to matter who sets the goal. Provided the goal is accepted and that at least partial success can be achieved and rewarded, the more challenging the goal, the more performance will be improved. On the other hand, they do say that goal setting will not work without feedback:

> The goal or target is practically useless if there is not enough information to keep performance on track. (*ibid.*, p66)

The precise theoretical interaction between goal setting and feedback is hard to disentangle. The provision of feedback may lead to spontaneous, implicit goal setting or, equally, the setting of goals will often lead to self-generated feedback.

Locke and Latham (1990) found in a review of 33 studies in which the combination of goal-setting and feedback had been compared with either alone that, despite the difficulties of isolating the two effects, and a variety of methodological flaws, the results were 'remarkably consistent ..... neither is really effective without the other' (p197). They went on to explain the different roles of the two:

> Feedback tells people what is; goals tell them what is desirable. Feedback involves information; goals involve evaluation. Goals inform individuals as to what type or level of performance is to be attained so that they can direct and evaluate their actions and efforts accordingly. Feedback allows them to set reasonable goals and to track their performance in relation to their goals, so that adjustments in effort, direction, and even strategy can be made as needed. Goals and feedback can be considered as a paradigm case of the joint effect of motivation and cognition controlling action. (p197)

Nevertheless, in practice, the precise cause is not important; the field interventions described by Locke and Latham suggest that the combination of feedback and goal setting can be extremely effective in improving performance. Given the wide variety of contexts in which this result has been found (Locke and Latham, 1990; Locke *et al.*, 1981), it seems likely that it will transfer to teaching.

Focus: 'ego-' or 'task-' involving

In the literature on motivation theory, a number of different types of motivation are identified, the basic distinction being between intrinsic and extrinsic motivation. These are defined by Deci *et al.* (1991) as follows (although in practice it is acknowledged that the distinction is not always clear: see Cameron and Pierce, 1994):

> Intrinsically motivated behaviors are engaged in for their own sake - for the pleasure and satisfaction derived from their performance. When intrinsically motivated, people engage in activities that interest them, and they do so freely, with a full sense of volition and without the necessity of material rewards or constraints ...
>
> Extrinsically motivated behaviors, on the other hand, are instrumental in nature. They are performed not out of interest but because they are believed to be instrumental to some separable consequence. (Deci *et al.* 1991, p328)

This division has been refined in the light of evidence that extrinsically motivated behaviours differ in the extent to which they represent self-determined as opposed to controlled responses, and the extent to which constraints have been internalised. Deci and Ryan (1985; Deci *et al.*, 1991) therefore divide extrinsic motivation into four types: *external regulation*, in which the incentive is wholly external; *introjected regulation*, in which an external control has been internalised as a feeling such as guilt; *identified regulation*, in which a person has come to value the behaviour and identified and accepted a formerly external control, and therefore feels a sense of choice about that behaviour; and *integrated regulation*, in which the regulatory process is fully integrated with the individual's sense of self, and behaviours are fully self-determined. The latter differs from intrinsic motivation in that an activity will be seen as important because of a valued outcome, rather than for interest in the activity itself. For Deci and Ryan the important distinction is between autonomous forms of motivation (i.e. intrinsic motivation, identified or integrated regulation) and non-autonomous forms (external or introjected regulation or 'amotivated' – synonymous with learned helplessness) (see Fortier *et al.*, 1995 p259). This theory has been heavily criticised by Locke and Latham (1990), who describe it as 'convoluted and constantly changing' and 'not well enough developed and articulated to make predictions possible'.

Maehr (1983, p193) makes a slightly different distinction, which stresses the goals of motivated behaviours. He divides 'intrinsic' motivation into 'task-involved' (concerned with mastery) and 'ego-involved' (concerned with beating others). Similarly 'extrinsic' motivation is split into 'social solidarity' (gaining approval) and 'extrinsic' (obtaining reward). These four types are seen as forming a continuum. There is some research evidence that the balance between task- and ego-involvement can be manipulated by changing the way feedback is presented (Nicholls, 1983; Butler, 1988). Nicholls summarises the evidence:

> Ego-involvement is likely to predominate over task-involvement when conditions, such as competition, induce self-focus or self-evaluation. Ego-involvement more than task-involvement implies evaluation of one's capacity compared to that of others, self-awareness, and perception of learning as a means to an end. (Nicholls, 1983, p215)

And the resultant effects on performance:

> ... children who perceive their ability (compared to that of others) as unacceptably low are not likely to learn effectively when they are ego-involved. When, on the other hand, children are task-involved, their capacity relative to that of others is not a concern. Instead they will focus their attention on the business of learning. Accordingly, their learning will not be impaired. (*ibid.*, p216)

In Butler's (1988) study, three kinds of feedback were given: 'task-involving', which consisted of comments on the work and how it could be improved, 'ego-involving', which consisted of normative grades, and to a third group, both types were given together. Performance was measured on two tasks, one of convergent and one of divergent thinking. Task-involving feedback was found to maintain interest and task involvement and to improve performance considerably on both convergent and divergent thinking tests. Ego-involving feedback, on the other hand, maintained immediate interest and performance in convergent thinking only for high achieving students; both were undermined for low achieving students, as were immediate divergent thinking and subsequent interest and performance on both tasks for both levels of achievement. When both types of feedback were given, the effect was largely the same as for grades only.

This general finding is confirmed by Kluger and DeNisi (1996): 'FI cues that seem to direct attention to task-motivation or task-learning processes augment FI effects on performance', although there is no attempt to quantify this. They explain

this phenomenon in their *Feedback Intervention Theory* by saying that certain types of feedback may direct attention away from the task and cause the person instead to focus on goals of the self. Feedback which draws attention to comparisons between individuals (normative feedback such as grades), or which makes an individual aware of how they are perceived (personal feedback, as opposed to feedback from a computer), or which directs attention towards self-esteem (discouragement or praise) are all examples of this.

Given the number of studies which focus on how feedback affects motivation or interest in a task, it must be said that the relationship between interest or motivation and performance is not always clear, and a number of studies have reported apparently counter-intuitive results (see Butler, 1988). Nevertheless, Deci *et al.* (1991, p331) provide substantial evidence that academic performance and autonomous forms of motivation are highly associated. Fortier *et al.* (1995 p261) go further, asserting that the relationship is causal, and that autonomous motivation produces higher creativity, less dropout, more cognitive engagement and better conceptual learning. Their evidence that the form of motivation can be influenced and that it causes improved performance is largely derived from laboratory research, so its transfer to field settings is unproven.

To summarise, it seems that feedback that conveys the same information can be manipulated to focus on either the task or on the individual's performance relative to others. In general a task focus produces better performance, particularly for lower achievers.

Focus: 'norm-referenced' or 'self-referenced'

A slightly different distinction is made in other studies, although this is somewhat obscured by unfortunate terminology. McColskey and Leary (1985) provided two different types of feedback: 'norm-referenced', which compared an individual's performance to that of others (i.e. what Butler and others have called 'ego-involving'), and 'self-referenced', which compared an individual's performance with other measures of their ability. They found that, for feedback that conveyed the message of failure (negative feedback), if it was norm-referenced it led to lower self-esteem, expectations and motivation, while if it was self-referenced it produced

increased attribution to effort and higher expectations for future performance. However, when the feedback signified success (positive feedback), the effect seemed to be reversed.

An experiment reported by Slavin (1980) to test the effect of self-referenced feedback found it could improve performance. Students were given points for improving on their past performance and improved significantly more (effect size 0.42) than the control group who received only the traditional grades. This applied equally across the ability range as measured by the pre-test.

These findings support somewhat mixed conclusions, although once again it seems that feedback that compares one's performance to others' is likely to be detrimental if performance is poor. Feedback which focuses on an individual's performance relative to their past achievements may lead to improved performance.

Perception of receiver: 'informational' or 'controlling'

A number of studies (e.g. Lepper *et al.*, 1973; see Deci and Ryan, 1992) have reported an *overjustification effect*: the tendency of additional extrinsic rewards to reduce intrinsic motivation. This is explained in terms of attribution theory, according to which in the presence of an extrinsic reward people attribute their behaviour to that, despite having been previously intrinsically motivated. When the external reward is removed, future motivation and performance decrease.

An alternative explanation of this phenomenon is provided by the *cognitive evaluation theory* of Deci and Ryan (1985), which states that an individual's level of autonomous motivation in a given situation depends on that person's feelings of competence and self-determination. Thus the effect of an extrinsic reward could depend on how it is perceived: if it is seen as 'informational' it could increase feelings of competence and increase motivation; if it is seen as controlling it would reduce intrinsic motivation. Reward which is contingent on participation is likely to be seen as controlling; reward contingent on performance may be seen as informational if a person performs well, or as controlling if they perform badly.

Providing teachers with feedback on their performance could certainly be perceived as introducing an extrinsic motivational factor, and thereby causing an 'overjustification' effect. Equally, feedback might be perceived as 'controlling' and

so undermine intrinsic motivation. One of the findings of Brinko's (1990) review was that 'sources of feedback should be lower or equal in status to the recipient' (p3). Deci and Ryan (1987) quote a number of studies which have found that surveillance (e.g. by video camera) or the prospect of evaluation (even if positive) reduces intrinsic motivation (see also Boggiano and Pittman, 1992; Deci and Ryan, 1992; Deci *et al.*, 1991).

However, a meta-analysis by Cameron and Pierce (1994) of studies that have examined the effects of extrinsic reward on intrinsic motivation shows a rather more complex picture. They found that the effect depends on the type of reward given (whether verbal or tangible), the expectancy of the reward (whether it is expected or not) and the contingency of the reward (whether it is dependent on the individual's performance or simply given for participation) as well as on the way intrinsic motivation is measured. Four measures were used: attitude, time freely spent on the task after removal of the reward, performance during the free-time period, and the subject's willingness to volunteer for future studies without reward. Cameron and Pierce found that '...in the laboratory, overall, reward does not negatively impact intrinsic motivation on any of the four measures'. Indeed, in some combinations of the above variables – for example, when the reward was verbal or when it was contingent on performance – motivation was significantly increased. They conclude,

> Rewards are detrimental only under a highly specified set of circumstances. That is, when subjects are offered a tangible reward (expected) that is delivered regardless of level of performance, they spend less time on task than control subjects once the reward is removed. The same condition has no effect on attitude. (p395)

The implications of this for teaching are not clear, but it does seem that feedback that supports feelings of autonomy and control is more likely to improve performance than feedback that is seen as controlling.

Valence or sign: positive or negative feedback

In this context, positive feedback is that which indicates a high level of performance or 'success', and, by implication, conveys the value judgement of approval; negative feedback, on the other hand, indicates a low level of performance or 'failure', with the implication of being unsatisfactory.

According to a behaviourist view, learning to improve one's performance in response to feedback would be seen as a form of *operant conditioning*. Positive feedback would therefore be a *reinforcer* and would be expected to lead to an increased tendency to repeat successful behaviour and thus to perform better. This view has diminished in currency over recent years and has been superseded by a cognitive *information-processing* view, in which the role of feedback as correction is stressed (Mory, 1992; Bangert-Drowns *et al.*, 1991).

A widely held example of the latter is Deci and Ryan's (1985) *cognitive evaluation theory*, according to which the perception of competence is motivating, provided it is in a context of self-determination. Deci and Ryan (1987, 1992) quote a number of studies to show that,

> positive feedback tends to increase intrinsic motivation, presumably because it enhances people's experience of competence. (1992, p13)

Harackiewicz *et al.* (1992) summarise the research evidence:

> A variety of competence cues were shown to both enhance and undermine intrinsic motivation through the processes of perceived competence, competence valuation, performance anxiety, and perceived control. ... Cues that lead individuals to perceive themselves as competent, or to value competence, may have positive effects. However, competence cues also may undermine perceptions of personal control, and can arouse performance anxiety, both of which have negative implications for subsequent interest. (p133, 134)

A further theoretical orientation which predicts differential effects for positive and negative feedback is *self-efficacy theory* (Bandura, 1986). According to this, positive feedback leads to an increase in self-efficacy, defined as 'people's judgements of their capabilities to organize and execute courses of action required to attain designated types of performances' (p391), which in turn leads to raised self-set goals and a resultant improvement in performance.

The sign of feedback may also influence a person's attribution for their performance. Weiner (1992, p279) suggests that success is more likely to be attributed to internal causes (such as ability or effort), while failure is more likely to be accounted for by external causes (e.g. luck or the difficulty of the task). However, it is the stability of the attributed cause that determines a person's reaction to that feedback and the effect on their subsequent performance. Attribution to stable causes

(such as ability or task difficulty) is more likely to lead to feelings of helplessness and to worse performance than attribution to unstable causes (effort or luck).

An alternative to the broad consensus of these views which see positive feedback as generally more beneficial than negative is put forward by some writers. Mesch *et al.* (1994) found that negative feedback led to higher goals and better performance on a simple task, and Podsakoff and Farh (1989) report that positive feedback can produce complacency. Waldersee and Luthans (1994) found that positive feedback had a debilitating effect on the performance of routine tasks. They account for this in terms of a *control theory* (Podsakoff and Farh, 1989) of behaviour in which negative feedback is perceived as a spur to action in order to remedy a deficiency, whereas positive feedback signals that there is no need to change. In the case of 'habit controlled behaviors' (i.e. highly routine tasks in which little or no conscious control is required to perform them), positive feedback may simply be a disruption. However, Mesch *et al.* (1994) point out that the long term effects of negative feedback are unknown but could include feelings of learned helplessness and loss of self-esteem, and thus ultimately reduced performance.

In a review of the literature on feedback effects, Ilgen *et al.* (1979) found that negative feedback is frequently misperceived and therefore likely to have less effect. Finally, Brinko (1990) gives the following advice, which is aimed specifically at suppliers of feedback to teachers: 'provide a generous amount of positive feedback with limited and carefully selected negative feedback', and suggests that the negative should be sandwiched between positive feedback. However, in this she does appear to be going beyond the evidence of the studies reviewed, and the effectiveness of this recipe must be judged as unproven.

These apparently conflicting results are hard to reconcile. On the one hand, positive feedback may be seen as reinforcing successful behaviour or as generating feelings of competence which have been shown to increase motivation and therefore enhance performance. On the other hand, in some situations, negative feedback is likely to lead to the setting of higher goals, which has been shown to lead to improved performance. However, in all cases is it not simply the sign of the feedback that is important but the way it is appraised and processed in relation to the individual's goals (Locke and Latham, 1990).

Timing: immediate or delayed

The common sense position that immediate feedback would be expected to have most effect is endorsed in reviews by Ammons (1956) and Brinko (1990). Once again, however, the issue is far from simple. In a meta-analysis of the effect of timing of feedback on verbal learning, Kulik and Kulik (1988, p79) state that 'In spite of the vast amount of attention given to feedback timing over the years, researchers still disagree about its importance in human learning'.

Some studies have shown that 'learners performed even better on a retention task when feedback was delayed', especially on higher level cognitive tasks (Mory, 1992, p6). This phenomenon has been called the *delay-retention effect* (DRE). However, other studies have failed to find the DRE, and Kulik and Kulik (1988) report that in 'applied' (i.e. classroom) studies, the mean differential effect size between immediate and delayed feedback is just 0.28. They conclude that delay of feedback is beneficial only under specific and somewhat artificial conditions; for conventional educational purposes, immediate feedback is preferable. One further meta-analysis, this time on the effects of feedback in computer-based instruction (Azevedo and Bernard, 1995), has examined the effect of timing of feedback and the post-test. They conclude: 'It is clear from these results that immediate delivery of a feedback message provides the best instructional advantage to the student' (p15).

The advice given by Brinko (1990) is that positive feedback should be given immediately, but negative feedback, because it is less accurately perceived and more likely to be forgotten, should be given just before the next performance. She also advises that frequent and repeated feedback is more effective, but reports the finding of Ilgen *et al.* (1979) that too frequent feedback can have an adverse effect.

The possibility of drawing any conclusions about the long term effects of feedback is seriously hampered by the tendency of all studies to be of short duration. For example, Waldersee and Luthans (1994) found that corrective (i.e. negative) feedback improved performance as compared with positive feedback and control groups, but 'satisfaction with supervision declined significantly for both the positive and corrective feedback groups.' (p93). We are left to wonder whether, had the study continued for more than three weeks, the effects of this dissatisfaction would have resulted in a lowering of performance. In the meta-analysis by Bangert-Drowns *et al.*

(1991), only 9 of the 58 effect sizes calculated came from studies lasting more than two weeks, and, of the remainder, 37 lasted just one week.

Lyakowski and Walberg (1982), in a meta-analysis of a number of instructional effects (including 'corrective feedback'), report that, despite finding large effect sizes (overall mean effect size 0.97), 'It cannot, of course, be concluded that such results can be sustained over long time periods; additional longitudinal research is needed on this question' (p570).

In trying to account for the prevalence of the misperception that feedback always improves performance, Kluger and DeNisi (1996) point to the confusion between the feeling that feedback is psychologically reassuring and desirable and the question of whether it benefits performance. However, they speculate that in the long term, by increasing satisfaction in the task, feedback may indeed increase 'long range persistence' and thus lead to improved performance.

An interesting feature of the research literature is illustrated by the surprising number of studies in Bangert-Drowns *et al.*'s (1991) meta-analysis which seem to have allowed the feedback to be used even before students had formed their own answers. An analysis of the answers suggested that these students were indeed copying from the 'feedback'. These studies were coded as 'uncontrolled for presearch availability' (i.e. so-called 'feedback' about the correct answers was available to students) and were found to produce slightly negative effects of 'feedback' on performance.

Thus the evidence generally supports giving immediate rather than delayed feedback in order to have maximum effect on performance. However, this conclusion must be somewhat tentatively applied to the context of giving performance feedback to teachers, since the evidence is mixed and drawn from a limited range of situations.

Specificity: general or focused

A number of studies have tried to compare the effects of giving different amounts or kinds of information in their feedback. However, as Mory (1992) observes, 'studies that have examined the question of the type and information which should be included in feedback have not yielded very consistent results'. (p12) Bangert-Drowns *et al.* (1991) report a clear finding that 'corrective' feedback (i.e.

feedback which provides either the correct response or an explanation, or requires the respondent to repeat until correct) is more effective than simple right/wrong feedback. Indeed, of the studies which used corrective feedback and also 'controlled for presearch availability', the mean effect size was 0.58 (n=30). These two factors accounted for virtually all the negative effects found in that study: the experiments where feedback had a detrimental effect on performance were almost all ones in which either the feedback gave no information about the correct answer or where so-called 'feedback' was available to the respondents before they gave their initial answers.

Another question addressed by some studies is whether feedback should be given to individuals or to a group. For example, Archer-Kath *et al.* (1994) state that 'for feedback to have maximal impact, it needs to be focused on the actions of individual group members (not the group as a whole)' (p693).

Brinko (1990) provides a list of conditions which 'tend to make feedback more effective', but her definition of 'effective' is broad and unclear and seems concerned more with changing behaviour than with necessarily improving performance. Nevertheless, her advice is that

> Feedback should contain concrete information .... Feedback should contain specific data .... Feedback should be focused .... The content of the feedback should reduce uncertainty for the recipient .... The content of the feedback must be relevant and meaningful to the recipient .... The content of the feedback must relate to goals which are defined by the recipient.

In general, it seems therefore that specific feedback that gives individuals direct and relevant information about their task performance is most likely to bring about improvement.

Credibility

Brinko (1990) advises that 'feedback should contain accurate data and irrefutable evidence', and clearly the amount of faith that recipients have in any feedback they get would be expected to influence how they are affected by it. Ilgen *et al.* (1979) report that the more credible a piece of feedback is seen to be, the more likely it is to be perceived accurately, and that credibility – and therefore acceptance – depends on characteristics of the source of the feedback such as expertise, reliability and trust. Feedback from multiple sources is generally seen as more valid and

reliable, and therefore more likely to be effective (Brinko, 1990, 1993). The most credible single source of feedback is the self: 'self-generated feedback was more credible than feedback from the organization or superior and significantly increased performance' (Brinko, 1990, p3; see also Ilgen *et al.*, 1979).

It seems to be widely found that positive feedback is more likely to be perceived as valid than negative (Jussim *et al.*, 1995; Moreland and Sweeney 1994). Ilgen *et al.*'s (1979) review stresses the importance of acceptance of the feedback as a factor mediating its effect, and reports that 'negative feedback was accepted only if it came from a high status source'.

Credibility therefore seems to be an important factor in determining the effect of feedback on performance. The more credible feedback is, the more likely it is that it will improve performance.

Significant variables 3: Individual characteristics of the receiver

Level of involvement

The prevalence of the use of laboratory experiments to determine feedback effects gives rise to a large number of studies in which the tasks set are unlikely to involve the subjects to any extent. Moreover, even in field settings, the tasks may be inconsequential. This is illustrated by Bangert-Drowns *et al.*'s (1991) meta-analysis of the effects of feedback on test performance, in which only 8 out of 58 effect sizes came from measures that counted for test grades. In this context, it may be fair to question findings such as Mesch *et al.*'s (1994) conclusion that corrective feedback (i.e. feedback emphasising deficiencies in performance) improves performance more than positive feedback (See Waldersee and Luthans, 1994, for a discussion of this issue).

It may also be worth noting that a surprisingly large number of experiments reported in psychology journals are performed on American psychology undergraduates. While it may not be fair to claim that such people are in any way abnormal (!), it does seem reasonable to question whether they are representative of other groups to whom the results of this research are presumed to transfer. In

reflecting on the limited generalisibility of the available theory, Maehr (1983, p190) observes that 'much of achievement motivation theory may be limited to the roles of white, middle-class males'.

It should also be stressed that achievement motivation is by no means the only driving force behind teachers' behaviour. One would expect that for many teachers, values such as altruism and the desire to help their students could well be more salient than their own personal desire for achievement. In real life, motivation is likely to be much more complex than in trivial experiments, and the level of involvement with a task may well be an important factor in mediating the effects of feedback on task performance. However, this interaction as yet awaits research attention.

Self-efficacy and self-esteem

Differential effects of feedback on performance between those high and low in self-esteem have been found in a number of studies. Ilgen *et al.* (1979) report that following positive feedback, individuals high in self esteem improved their performance more than those of low self-esteem. Conversely, following negative feedback, the performance of those low in self-esteem decreased more. They also found that levels of self-esteem were associated with differences in acceptance of different sources of feedback: those of high self-esteem were more likely to rely on self perceptions, while for those low in self-esteem, external sources of feedback were more salient. Butler's (1988) experiment has already been mentioned: ego-involving feedback (normative grades) was found to have a more negative effect on performance for those low in achievement than for high achievers. However, it must be stressed that self-esteem is not the same as achievement, and individuals high in one may not necessarily be high in the other.

Nicholls (1983) reviews the evidence on task- and ego-involving feedback and concludes that:

> Many studies show that individuals with low perceived ability perform better in task-involving than in ego-involving conditions and at a similar level to those with high perceived ability. Those with high perceived ability perform at similar levels in both states. (p217)

A possible explanation for this phenomenon is given by Kluger and DeNisi (1996):

> ...negative FI is more likely to direct attention to the self among participants low in self-esteem than among those high in self-esteem, but positive FI may have the opposite effect. (p269)

Alternatively, this may be partially accounted for in terms of the different attributions people have for their performance in each of the two states, as described above. Attribution of failure to stable causes has been shown to lead to much lower expectations of future success, and individuals high in self-esteem are more likely to attribute success to ability (Weiner, 1992, p261). However, that attribution can be altered in order to increase persistence and performance (Forsterling, 1985). One further explanation comes from the theory of *goal-setting* (Locke and Latham, 1990) according to which self-efficacy influences the level of difficulty of goals set and the individual's commitment to them. Both these factors have been shown to affect task performance.

Jussim *et al.* (1995) provide a theoretical framework for analysing reactions to feedback, which are seen as a product of three competing tendencies. The first is 'self enhancement', the desire to see one's self favourably. This predilection accounts for the fact that positive feedback is generally seen as more accurate and is more likely to be attributed to internal causes (see also Moreland and Sweeney, 1994). Secondly, 'self consistency', the propensity to assimilate ideas if they are consistent with past experience, and reject them if they are inconsistent. From this it follows that for those high in self-esteem, positive feedback will be seen as more accurate and will lead the receiver to take more responsibility for her or his performance. Conversely, for those low in self-esteem, the same would be true of negative feedback: it is perceived as more accurate and more likely to be attributed to internal causes. According to a 'strong' version of self-consistency theory, those low in self-esteem will actually feel better on receiving negative feedback. The consequences of these differential attributions are that after failure, individuals low in self-esteem are more likely to feel incompetent and consequently have lower expectancies of future success, lower motivation and poorer performance (Weiner 1992, p261, Anderson & Jennings, 1980). The third tendency is 'accuracy', the desire to evaluate one's performance and abilities accurately.

There is some debate in the literature as to the level of generality of the construct of self-efficacy or self-esteem that is most relevant. Self-efficacy is usually

a judgement made in relation to a specific task and context, although it may vary in generality. Self-esteem, on the other hand, is a more global and stable self-judgement of a person's capability. In an experiment to study people's reactions to feedback, Jussim *et al.* (1995) found that 'global self-esteem influenced reactions to feedback substantially more than did specific expectations' (p353). However, global self-esteem is not easily influenced, and is 'extremely stable over short periods of time' (p333) and therefore not affected by feedback. A similar conclusion was drawn from a correlational field study by Moreland and Sweeney (1994):

> Although GSEs [global self-expectancies, equated with self-esteem] are usually less task relevant, they may also tend to be clearer and/or more stable and therefore more likely to influence a person's reactions to a performance evaluation. (p172)

However, they do point out that this may also be because their measurement of GSE was more reliable and valid.

To summarise, it can be said that an individual's level of self-esteem may well affect how they respond to feedback. Those high in self-esteem are probably generally more likely to improve in response to feedback, and feedback that can have a detrimental effect (e.g. feedback that focuses on ego-involvement or negative feedback) is likely to be worse for those low in self esteem. However, the numerous interactions between self-esteem and other variables, such as the sign of the feedback, its focus on the self or the task, the causal attributions made by the recipient, make the situation rather complex. In addition, a number of other similar but distinct constructs (level of achievement, achievement orientation and specific performance self-efficacy, for example) may be compounded with self-esteem and may be more relevant in accounting for differential effects of feedback on performance.

Attributions for success and failure

Forsterling (1985) reviews studies of *reattribution training* which have tried to alter individuals' causal attributions for their behavioural outcomes. These studies have adopted one of two slightly different approaches, depending on their theoretical perspective. If the perspective is from the theory of *learned helplessness* (Abramson, Seligman and Teasdale, 1978), then the objective is to increase the individual's feelings of control over the outcomes. In this case, attributions of both success and

failure to effort are felt to be most desirable. However, if the underlying orientation is guided by either *attribution theory* (Weiner, 1992) or *self-efficacy theory* (Bandura, 1986), then attributing success to high effort could imply a lack of ability – especially if the task is not particularly difficult – which would be expected to lead to lower future performance. Hence it is most desirable to attribute success to ability and failure to lack of effort, or luck. Forsterling (p509) concludes that '... attributional retraining methods have been consistently successful in increasing persistence and performance'.

An interesting variation is found in a study by Anderson and Jennings (1980) in which subjects were persuaded to attribute their failure not to effort but to their particular choice of strategies. As well as increasing their expectations of future success, this also led them to focus more on their strategies, monitoring and modifying them in response to their failure and thereby learning from experience in a way that the control group did not.

These different perspectives are not necessarily mutually exclusive, given that an individual may not attribute success or failure to a single cause. Feedback that encourages people generally to view their level of effort as determining the outcome they achieve, and to attribute success to their ability and failure to the specific strategies used is likely to lead to improved performance.

Locus of control

The construct of *locus of control* was originally proposed by Rotter (1966) to account for the differences between individuals in their expectations about the relationship between their own behaviour and the reinforcement (i.e. reward or punishment) they receive. Those with an *internal* locus of control believe the two are reliably related; those whose locus is *external* believe they are unrelated. The original theory has been extended and modified by others. For example, DeCharms (1968) has defined *locus of causality* in terms of the amount of freedom individuals ascribe to their behaviour. Weiner (1992) uses the terms *locus of control* and *locus of causality* interchangeably to refer to the dimension which distinguishes attributions for events between internal or external causes.

Ilgen *et al.* (1979) show that a number of studies with subjects from a variety of groups have found that individuals with an internal locus of control out-performed those with an external locus when task-supplied feedback was the only kind available. The reverse was true when feedback was available only from the experimenter. Those with an internal locus were also more likely to accept or believe the feedback they received.

The implications of this for maximising the impact of performance feedback on teachers are not clear. Those with an internal locus of control are apparently more accepting of feedback and more likely to believe that they are able to influence outcomes. These two factors make it a reasonable conjecture that their performance will be more likely to improve in response to feedback than the performance of those with an external locus of control. However, this conjecture can only be somewhat tentative at this stage.

Achievement orientation

The foundation for much of the theory of achievement motivation can be found in Atkinson and McClelland's *expectancy-value theory* (Atkinson and Feather, 1966; McClelland, 1961). According to this theory, individuals differ in the predominance they show between the motive to achieve success and the motive to avoid failure. Thus an individual's tendency to attempt a particular task is a product of their motive to achieve, their expectancy of success and the value they place on that success. Likewise, the tendency to avoid a task is the product of the motive to avoid failure, the perceived likelihood of failure and the unattractiveness of the failure.

Harackiewicz *et al.* (1987) and Harackiewicz *et al.* (1992) describe the characteristics of achievement-oriented individuals. They are said to desire objective ability feedback, to show strong interest in diagnostic ability assessment, to become involved in activities that afford self-evaluation, to care more about doing well when performance is being evaluated, to hold high expectations for their performance and to show higher intrinsic motivation when feedback is positive. Individuals who are not achievement-oriented, on the other hand, tend to avoid ability assessment, to have lower performance expectations, to value competence less and are more likely to become anxious in evaluative situations. Harackiewicz *et al.* (1987) found differences

between the two types in their responses to feedback, particularly when the feedback had a normative, as opposed to task, focus. In this condition, achievement-oriented individuals enjoyed the task more and showed more interest in it than those who were not achievement-oriented.

Harackiewicz *et al.* (1992) propose a model to explain how 'competence cues' (i.e. feedback or expectations about an individual's competence in a particular performance situation) affect intrinsic motivation. They see the level of intrinsic motivation as determined by four factors: an individual's *perceived competence*, *competence valuation*, *performance anxiety*, and *perceived control*. These factors are also partly a product of the individual's achievement orientation, as described above. However, although in experiments achievement orientation seemed to determine prior performance expectations and reactions to different kinds of feedback, the actual performance levels of the two groups were experimentally constrained to the same standard. Hence it is not clear what effect differences in achievement orientation may have on task performance.

However, there is evidence that goal-setting improves the performance of those with high achievement needs most (Harackiewicz *et al.*, 1992, p131), and that achievement orientation is associated with the tendency to attribute performance to effort and consequent higher persistence in the face of failure (Weiner, 1972). Given these findings, and the characteristics associated with achievement orientation described above, it seems likely that those who are high on this measure will improve their performance more in response to feedback. Once again, though, this conclusion is very speculative.

Receptiveness

It seems plausible that the recipients of feedback are more likely to be affected by it if they are initially receptive to getting it. Once again, Brinko (1990) has some advice, culled from the literature: 'Recipients of feedback should be volunteers or at least receptive to the process ... The recipient should be able to select the mode of feedback.'

Adequacy of original performance

In studies of the effects of feedback on tasks requiring relatively low level learning, it is not surprising that feedback (knowledge of results) has more effect when the original response was wrong. Bangert-Drowns *et al.* (1991) report an appreciable correlation (0.48) between the rate of errors made by students during instruction and the effect size of feedback on performance. They interpret this by saying that the role of feedback is largely corrective. This corrective role is also stressed by Azevedo and Bernard (1995), who found a weighted mean effect size of 0.80:

> Feedback has to be regarded as one of the most critical components of computer-based instruction, its objective being to provide students with appropriate responses thus allowing them to rectify learning impasses. (p13)

A similarly large effect is reported by Lyakowski and Walberg (1982). Their meta-analysis of the effect of 'corrective feedback' on learning outcomes calculated a mean effect size (unweighted) of 0.94. However, their definition of 'corrective feedback' is broad, and includes any form of testing 'whether oral, written, or practical problem solving' (p561). They also found quite a wide range of effect sizes and apparently included some large ones from studies with 'inadequate generalizability', so the true effect may be somewhat smaller.

The effect of feedback depends not only on the accuracy of a response but on the confidence with which it is given, or *response certitude*. Mory (1992) states that feedback gets most attention and is most effective at error correction when the answer is wrong but response certitude is high.

Whether these results will transfer to more complex tasks, and to situations where feedback gives an indication of performance rather than simply correcting errors, is hard to say. Nevertheless, large effect sizes have been found in a range of field settings and it is at least plausible that feedback that aims to correct specific errors or inadequacies in teaching will have a similar effect.

3.3 DISCUSSION AND CONCLUSIONS FROM THE REVIEW

Difficulties of applying these results to improve teaching

Most of the theories which try to account for the effects of feedback on performance are limited in scope, too vague to be readily operationalised or to enable predictions to be made, and/or supported by only some of the available evidence. The validity of many of the concepts used – for example, motivation, locus of control, self-esteem – becomes very questionable when they are used as global attributes of a person, rather than being seen as context-dependent (Leo and Galloway, 1996). As a result, the role of theory in advancing knowledge in this area is extremely problematic. There is a need for research that is 'condition-seeking' rather than 'theory-testing' (Greenwald *et al.*, 1986).

Given these limitations, transferring inferences from one context to another must be done with some caution. In particular, the transfer from laboratory experiment to classroom involves a big jump, and much of the literature is characterised by unacknowledged leaps of this kind. Predictions of what will happen in a certain situation must be based on evidence from either the same situation or from a large number of studies in essentially similar contexts. There is a great need for experimental studies in field settings.

A further limitation is the devastating absence of long-term studies. It is important to know the long-term effects of providing people with feedback, and only experiments conducted over a long period will establish this.

It will always be difficult to isolate the effects of feedback, and many of the variables that appear to mediate its effects are highly interrelated. Although explicit feedback can be controlled, in many performance situations implicit feedback on performance will be gained from the task itself. Similarly, it is hard to separate monitoring from feedback. For example, in Waldersee and Luthans's (1994) experiment, all groups, including the control, improved their performance: the employees knew they were being monitored and presumably performed better in response.

Finally, in the context of teaching it is by no means straightforward to define 'performance' in a way that would be either likely to gain broad agreement, or would be possible to measure satisfactorily. As has been argued in Chapter 2, student outcomes provide only a limited measure of teaching performance, and other aspects of performance are also hard to measure.

Tentative summary of the conditions under which giving feedback to teachers will have maximum impact

With all the above reservations in mind, there are nevertheless some apparently clear findings in the research literature. One can therefore conjecture that the following conditions maximise the likelihood that giving feedback to teachers will improve their performance. However, any sustained feedback effects presumably arise as a result of changes in the behaviour of the recipients, which are notoriously hard to influence in any significant way (Tymms, 1995), so we should perhaps not expect long term effects to be large. The factors are divided here according to whether they are likely to be fixed or alterable (Bloom, 1979).

Factors which are determined by the task and its context, or are apparently stable characteristics of the individuals taking part

- The task is simple in nature.
- The task is such that by trying harder they are likely to perform better.
- The feedback in question is the main or only source of information about performance (this may not maximise performance, but will maximise the effect which is specifically attributable to that particular feedback).
- The task is such that teachers are likely to feel involved in it and one in which their performance is likely to be important to them.
- The recipients of feedback have high self-esteem, an internal locus of control and are achievement-oriented.
- Feedback recipients are volunteers.

Factors which may be altered

- Individuals have clear, specific and challenging goals related to their task performance. Feedback provides information with which to measure performance against these goals. Even partial success can be rewarded.
- The feedback causes people to focus on the task, not on their performance relative to others.
- Feedback focuses on individuals' performance relative to their past achievements, rather than relative to others.
- The feedback is perceived as providing information and supporting self-determination, not as surveillance or control.
- The feedback generates feelings of competence.
- The feedback does not generate feelings of complacency.
- Feedback is given as soon as possible after performance.
- The recipients attribute their performance to their own efforts: they feel they have control over the outcomes.
- The recipients attribute their success to their ability or to the effort they have applied; they attribute failure to a lack of effort or to specific inadequacies, such as adopting a poor strategy.
- The feedback is perceived as being credible and accurate.
- Feedback is given to individuals on their individual performance.
- Feedback is specific and focused on the task.
- The feedback aims to correct errors or inadequacies.

Attempt at synthesis

In an attempt to synthesise and summarise the above conditions, the following general features emerge:

- Feedback can help to *focus* on particular aspects of a task, thus making them more salient and so increasing *motivation*, as well as helping to exclude extraneous aspects from *attention*. In certain cases, feedback may cause a person to focus on task-extraneous aspects such as their own feelings of inadequacy or lack of autonomy and will therefore not lead to improved performance.

- Feedback can have a *diagnostic* function, allowing people to see to what extent they are achieving their goals in different aspects of a task and so helping them to *account for* and *learn* from satisfactory outcomes and to *modify* less satisfactory ones.

In both these ways, feedback may lead to improvements in performance, provided those receiving it have clear and demanding task goals which they believe to be attainable and which they are already motivated to achieve.

Chapter 4

Overview of the Empirical Investigation

This chapter contains a brief summary of the aims and methodology of the empirical study. It also contains a note on the style in which the account has been presented and a methodological note on the use of significance testing.

4.1 OVERVIEW OF THE METHODOLOGY

The initial design for the study involved contacting a group of schools to ask for their participation in the project and collecting detailed quantitative and qualitative data from the teachers involved. A randomly selected half of them were to be supplied with certain kinds of feedback and the effects on their students' subsequent performance monitored. However, a number of factors caused this design to be modified slightly and the original sample of schools was augmented by an additional group. These two samples were used for two distinct experiments – albeit with similar aims – and are referred to here as Project 1 and Project 2 respectively.

The main reason for modifying the study in this way was the gradual attrition of the original sample when faced with the not inconsiderable demands of supplying the information requested at various stages. Each time a questionnaire or request for information was sent there was a significant proportion of non-response (or very delayed response), despite persistent reminders. The majority of teachers involved seemed to be very happy to participate and were often extremely apologetic for any inconvenience they had caused. However, it was clear that with the demands and

stresses of their work, some of which were alluded to in comments they made, the process of supplying me with information might well be a relatively low priority. As the sample dwindled, it also became clear from a detailed review of the research literature on the effects of feedback on performance (see Chapter 3) that the likely effect of the kinds of feedback I was intending to provide would not be large. Hence, in a small sample it would be hard to demonstrate that any difference between treatment and control groups was indeed a feedback effect and not just an accident of sampling (i.e. to achieve a 'statistically significant' result – see p75, below, for a discussion of this issue). Because of these concerns, it was decided to conduct a further experiment (i.e. Project 2) with a larger sample of institutions but requiring substantially less input from the teachers involved.

A brief outline of the specific aims and methodology of each of the two projects follows. A more detailed description of the process of data collection and analysis of the results of Project 1 can be found in Chapter 5 and Chapter 6, respectively. Similarly, Chapters 7 and 8 contain detailed accounts of the data collection and analysis of Project 2.

4.2 PROJECT 1

Description

An in-depth study of nine volunteer institutions to investigate teacher attitudes and responses to feedback.

Aims

1. To investigate the kinds of performance feedback being used by teachers, and their attitudes to and perceptions of that feedback.
2. To provide a group of teachers with feedback about the value added performance and attitudes of students they had taught and to measure any effects on:
    - teachers' self-reported behaviour;

- teachers' attitudes;
- teachers' self-perceptions;
- examination performance and attitudes of students subsequently taught by them.

3. To seek feedback from the teachers involved about their responses to the feedback sent and the kinds of feedback they would like to get.

Outline of methodology

Institutions in the sample were essentially volunteers. The findings of Project 1 could therefore not confidently be generalised to the population of all teachers. The fieldwork was conducted in eight stages:

1. Exploratory interviews

Exploratory interviews conducted by telephone to try to elicit comments about feedback used and attitudes towards ALIS.

2. Initial questionnaire

To collect data on attitudes (Likert items and open-ended questions) and uses of feedback. Also personal information. Responses used as a base-line (pre-test) measure.

3. Identification of teaching groups

Heads of department were asked to indicate which students were taught by which teacher(s). This information enabled teachers to be sent feedback specific to the groups they had taught.

4. Feedback

Feedback containing information about the intake characteristics, performance and attitudes of the students they had taught was sent to a randomly allocated half of the teachers.

5. Implementation-check questionnaire

Those who had received the feedback were asked how much time they had spent on it and how valuable and accessible they had found it.

6. Final questionnaire

To measure any changes in attitudes (using same Likert items as in initial questionnaire) and self-reported behaviour for both control and treatment groups.

7. Final interviews

To gain further insight into perceptions of the feedback and to validate the interpretation of attitudes from questionnaires.

8. Examination analysis

Student examination results were analysed to investigate differences between those taught by teachers who had had feedback and those who had not.

4.3 PROJECT 2

Description

An experiment providing selected departments in 192 randomly chosen institutions with different forms of feedback to investigate the effects on examination performance.

Aim

1. To measure any effects of each of three different forms of 'feedback' on the value added examination performance of the students in those departments.

Outline of methodology

The sample was chosen at random from all institutions registered in ALIS for both examination years 1996 and 1997. The English, French, Mathematics and Physics departments in each institution were randomly allocated to receive one of the following four treatments, such that each combination of subject and treatment occurred equally:

- *Departmental Information*. The Head of Department was sent a printout showing value added analysis of last years' results and target grades for this years';
- *Analysis by Teacher*. The Head of Department was sent the offer of the same analysis and targets, but on a class by class basis, if they returned class membership information;
- *TAMIS*. The Head of Department was sent a piece of software with which they could do their own analysis and target setting;
- *Control*. Nothing was sent.

Students' examination performance for those in each of the four groups was analysed to see whether any of the treatments had had an effect.

4.4 NOTES ON THE CONTENT OF THIS ACCOUNT

Critical and honest approach

In writing up this research, it has been the intention as far as possible to describe the whole process as it actually happened, rather than to present the type of idealised account typically found in journals and other research reports (Walford, 1991). It is hoped that, by presenting it in this way, the naïve but commonly portrayed notion of the objective, detached researcher is exposed as myth. Challenging this myth is a relatively rare approach, according to Walford, especially in accounts of quantitative research. Also, by adopting a consciously self-critical attitude to the methodology and

results, the genuinely complex and equivocal nature of social science research will be conveyed more honestly.

The danger, of course, is that methodology and results subjected to forceful and effective self-criticism might be mistaken for inadequate methodology. However, it is hoped that an honest account that acknowledges its limitations will be found more, rather than less, convincing. Nevertheless, the reader who is not used to seeing 'warts and all' is asked to guard against making unfair comparisons with more idealised reports.

On the use of tests of statistical significance

Since the time of Fisher, the use of significance testing in empirical social science has been widespread, if not obligatory. The notion that a particular result, found in a sample, could be just an accident of sampling rather than evidence of some characteristic of the parent population is one that must be taken seriously. However, significance testing as it is often practised – what Cohen (1994) has called 'mechanical dichotomous decisions around a sacred .05 criterion' – can be criticised on a number of counts. Briefly, these include the following:

Criticisms of significance testing

1. *It tells you the opposite of what you what to know*. Significance tests tell you the probability of getting a result as 'extreme' as you have, given the null hypothesis. What you want to know – and this is quite different – is how likely it is that the null hypothesis could be true, given the result you have just got. Cohen (1994) describes this as the 'inverse probability error' and both he and Carver (1978) give examples to illustrate the fallacy.
2. *It is logically nonsensical: the null hypothesis is always false*. It is impossible that in the population from which your sample is drawn the two means are exactly equal, or that the correlation is exactly zero. It is nonsense (and certainly not useful) to talk about the 'truth' of a null hypothesis which specifies a precise value (usually zero) for some population parameter when the only evidence about it comes from a sample (Cohen, 1994; Thompson, 1996).

3. *Its true/false dichotomy inappropriately stresses decision above inference*. In most research contexts (as opposed, say, to its use in quality control) it is not appropriate to have to make an all or nothing decision about whether to accept or reject a particular null hypothesis. It is absurd to have to have to conclude one thing if the result of an experiment gives p = 0.051 and the exact opposite if it were 0.049 (Eysenck, cited in Oakes, 1986, p26).
4. *It leaves out the most important information: the size of the effect*. It is not enough to know, as Tukey (1969) has said, 'if you pull on it, it gets longer'. Scientific advance requires an understanding of how much. Significance tests do not tell us how big the difference was, or how strongly related were two variables. Instead, they say more about how large our sample was (Thompson, 1992). A great deal more information can be extracted from an experiment if the focus is on parameter estimation, rather than hypothesis testing (Simon, 1974).
5. *It generates confusion between statistical and substantive significance*. The significance – in the true sense – of a result depends on the size of the effect found and whether it can be replicated. 'Significance' tests do not measure this, even imprecisely (Oakes, 1986), but are widely presented and interpreted as doing so.
6. *It is widely misunderstood*. Studies of practitioners' understanding of significance tests (e.g. Oakes, 1986) suggest that misconceptions (e.g. that a statistically significant result is highly likely to be replicated, or that the failure to reject a null hypothesis is evidence of its truth) are not sporadic but near universal. While this may not necessarily be the fault of significance tests, it is an argument against their use.
7. *It takes no account of any prior knowledge*. Even for the non-Bayesian, there are situations where the automatic output from significance testing must be tempered by prior knowledge (Oakes, 1986, p128; Carver, 1978, p392). Scientific advance proceeds by the accumulation of knowledge, not by results considered in isolation.
8. *It is open to easy abuse by selection*. The 'file drawer problem' (Rosenthal, 1979) refers to the over-representation in published work of statistically significant results, leading to overall bias. Research syntheses based on available studies are liable to over-estimate the size of an effect, because those that failed to

achieve statistically significant results are less likely to be published. Even within a study it is impossible to know how many 'non-significant' relationships have been tested, consciously or not, in order to find the 'significant' ones that are presented. The statistical significance of a result depends not just on the data, but on the way such findings were sought.

9. *It demands an unscientific asymmetry*. Carver (1978) describes the use of significance testing as a 'corrupt scientific method'. By considering the power of significance tests reported in social science journals, Cohen and others (see Cohen, 1990) have shown that the majority of studies published have a less than even chance of rejecting the null hypothesis, even where there is in fact a medium-sized effect. In other words, failure to reject the null hypothesis typically tells you absolutely nothing, other than that your sample was probably too small. Using such tests is as about fair as 'heads I win, tails we try again'.
10. *It puts unnecessary restrictions on sample size*. A large number of studies with small samples and similar results may provide more evidence about a phenomenon than a single large study, but taken individually none of them may have the power to achieve statistical significance. Even Fisher, who is often credited with much of the responsibility for the evils of significance testing, regarded the 5% level as arbitrary and took as a basis for knowledge the repeated finding of results at this level, rather than any single highly 'significant' result (Tukey, 1969). However, because of the orthodoxy of significance testing, these small studies may never be done, having been rejected at the planning stage as having insufficient power.
11. *It emphasises random errors at the expense of explanations*. Because significance tests, along with other forms of statistical analysis, enable us to side-step problems of inaccurately measured data (measurement error), and poor methodology (under-specified models) by aggregation with large samples, they may prevent us from adopting the ultimately more profitable strategy of addressing these inadequacies (Savitz, 1993).
12. *It requires a number of often unjustified assumptions*. The use of statistical tests of significance nearly always depends on making distributional assumptions about the statistic in question, and on the use of strictly random sampling. While distributional assumptions are sometimes acknowledged, and results may be

robust to their violation, the assumption of random sampling is often neither (Shaver, 1993). Also, precise 'p' values are highly sensitive to scale transformations and depend heavily on the (generally fairly arbitrary) choice of a particular measurement scale (Cliff, 1993, p497). Significance levels ('p' values) are often treated as far more accurate than is justified.

13. *It leads to wrong conclusions based on 'vote counting'*. Simply counting the number of studies that have found an effect and balancing them against those that have not is still a common component of many reviews. However, by ignoring the sizes of the effects and the samples, this 'vote counting' approach can lead to a conclusion opposite to that supported by the data considered as a whole (Hedges and Olkin, 1980).
14. *It perpetuates an adversarial tradition in social science*. On almost any issue studies can be found arguing for diametrically opposed conclusions, but a good many of the apparent differences are simply due to sampling variation (Hunter and Schmidt, 1996). Significance testing greatly exaggerates these differences, stressing individual results at the expense of an integrated overview of all the available evidence.

Alternatives to significance testing

A number of the critics of significance testing (e.g. Cohen, 1994; Thompson, 1996) make some suggestions of alternative ways of interpreting empirical results and allowing for their sampling variability. The following are based on these and other sources:

1. *Use better language*. The word 'significant' should not be used on its own when what is meant is 'statistically significant'. Better still, report that a particular null hypothesis was rejected.
2. *Look at the data*. Simple, flexible, informal and largely graphical techniques of exploratory data analysis, such as those described by Tukey (1977), aim to enable data to be interpreted without statistical tests of any kind.
3. *Report parameter estimates with confidence intervals*. A confidence interval contains all the information in a null hypothesis test, and more. Parameter estimates can often usefully be reported as standardised effect sizes.

4. *Replicate results*. Only by demonstrating it repeatedly can we guarantee that a particular phenomenon is a reliable finding and not just an accident of sampling. Internal replicability analyses such as cross-validation, the jackknife or bootstrap (Thompson, 1994) provide a means of assessing sample variability.
5. *Synthesise the results of multiple studies using meta-analysis*. This can provide an overview of findings in which the statistical significance of individual results has no part. Instead, results are pooled to give overall estimates of effect sizes and an understanding of the relationships among different variables.

In writing up this study, I have tried to follow these suggestions, wherever appropriate.

Chapter 5

Project 1: Data Collection

This chapter describes the methodology, instruments used and responses received in the collection of the data for Project 1. It is divided into eight sections according to the eight stages of the research outlined previously (see p72).

5.1 EXPLORATORY INTERVIEWS

The purpose of these interviews was to find out, without imposing predetermined structure or outcomes, what teachers' attitudes to feedback were and how they used it. It was also hoped that possible questionnaire items might be suggested by comments made in interviews. The format of the interviews was thus largely unstructured (Oppenheim 1992), leaving scope to follow up ideas as they arose.

The interviewees were originally volunteers attending a conference for users of ALIS and YELLIS in June 1996, and the interviews were conducted by telephone after the conference. Frey (1983) gives a number of advantages of using the telephone for surveys, which include high response rates and low interviewer influence on responses. However, the main advantage in this case was convenience, especially given the wide geographical spread of the volunteers.

I had hoped to interview ten teachers, but was unable to secure this many volunteers. Delays and difficulties in arranging the use of a telephone and recording equipment meant that lack of time also became a factor, and in the end only three interviews were conducted. I produced a loose schedule prior to conducting the

interviews which is reproduced in Appendix 5A. The interviews were audio recorded and the recordings transcribed (see Appendix 6A).

5.2 INITIAL QUESTIONNAIRE

Purpose

The initial intention in devising this questionnaire was to try to measure what might be expected to be significant covariates of any effect of feedback on students' performance. In other words, to measure certain characteristics of the respondents in order to be able to see what factors were associated with any feedback effects. Variables such as the sex and main subject of respondents, their reported uses of feedback, their reported attitudes to receiving feedback (in particular from ALIS), their attitudes towards ALIS (its perceived validity and value), and the stage of their development of using ALIS were all considered *a priori* to be possibly related to the way people would react to the feedback they received.

This last variable was included as a result of anecdotal and personal experience which suggested that familiarity with and effective use of the kinds of feedback provided by ALIS and this research were gained as a result of a learning process, possibly over many years. It was operationalised using an eight-point scale based on Hall & Loucks (1977), who argue that innovations go through a common series of stages in their adoption. These items formed what was conceptually similar to a Guttman scale, in that they were expected to be largely ordinal and cumulative: in other words, to represent a single dimension. However, none of the elaborate process of formal scale development was undertaken (McKennell, 1977; Oppenheim, 1992), and respondents were invited to select 'any of the following', so were not restricted to choosing only one item. The exact wording of this item can be found in the copy of the questionnaires used in Appendix 5B.

In addition to these variables, a further set of factors was derived from the literature on feedback effects (see Chapter 3). These factors included respondents' 'achievement orientation' (the degree to which they attach importance to their performance and value performance feedback), 'locus of control' (the extent to which

they perceive success or failure as within their control), and 'self-efficacy' (the extent to which they perceive themselves as effective teachers). All of these had been shown in the available research to be related to the effects of giving people feedback.

Finally, an attempt was made to gauge teachers' perceptions of the relative importance of different factors in influencing students' examination performance. This was done partly to see to what extent the intuitive impressions of those closest to the process of producing examination results (i.e. teachers) agreed with the research evidence from 'school effectiveness' (see Chapter 2). It was also thought that teachers might vary in the way they divided up the responsibility for examination performance, and that such variations might be related to other attitudes or effects. An innovative question format was used, with respondents being asked to divide a circle into sectors, each representing the relative importance of that factor (similar to a pie-chart).

The above variables, therefore, were included because of their possible mediation of the effects of feedback. However, at an early stage of the development of the questionnaire, an additional purpose emerged. It began to be clear that to expect a distinct and significant effect of giving extra feedback to be evident in one years' data was to be unrealistically optimistic. Given the size of the sample, time available, likely size of the feedback effect and the inherent instability of student performance, it seemed more likely that the examination results of the students of teachers who had had the feedback would be indistinguishable from the results of those whose teachers had had none. The main constraints – the size of the sample that could be worked with and the amount of time available – were largely determined by the scale of the project (i.e. a three year PhD with a single researcher) and could not really be overcome. The kinds of changes in teacher behaviour that would be likely to be manifested in improved student performance might well take a much longer time and a more significant intervention to become apparent (Hopkins and Lagerweij, 1996, p80-87). Nevertheless, it was felt that evidence for the beginnings of such a change might be seen in the form of changing attitudes and perceptions of those who had received the feedback. Thus, the focus of the project grew to include the attempt to measure the effects of feedback on the attitudes of those who received it. The initial questionnaire would therefore also have to serve as a pre-test measure of those attitudes.

Pilot version

A pilot version of the questionnaire was drafted and sent to 57 teachers in two schools whose ALIS co-ordinator had volunteered to take part at the ALIS/YELLIS conference. 15 replies were received, all from the same school.

Respondents were asked to comment on any questions they found 'unclear, meaningless or otherwise hard to answer.' Likert scale items (section B of the questionnaire - see Appendix 5C) were scored 'agree strongly' = 1 to 'disagree strongly' = 5, and the variance of the scores on each item was calculated as well as the correlation (Pearson product-moment) between each pair of items.

As a result of the responses to the pilot version, some changes were made to the 'personal details' section, adding new questions which seemed potentially useful ('sex' and asking whether they had taught an examination class in each of the years being studied) and pre-coding the answers to another ('subject taught'). A few of the Likert items were dropped or modified as a result of comments made or if a significant number of people had left them blank (items 14, 21, 26).[4] One item (item 6) was dropped because of very low variance of responses (everyone agreed with it); one (item 1) was modified to try to make it easier to disagree. Four tentative 'scales' were made by combining items which seemed to have common *prima facie* meaning and testing for internal consistency (Cronbach's alpha) and the strength of inter-correlations (See Chapter 6 for a full description of this process as applied to the revised version of the questionnaire). Several items were modified or dropped as a result of either not being correlated with any of the others (items 3, 12, 24, 28), or of not being correlated in the expected direction with other items in the same scale (item 10). Two other items (8 and 13) were modified slightly to try to make them clearer. Because of the small number and lack of representativeness of the returned questionnaires, any conclusions from this analysis were adopted very tentatively and the number and scope of the changes made was not great.

[4] The full version of the pilot questionnaire with the exact wording of all the items is reproduced in Appendix 5B (p243).

Summary of content of the revised questionnaire

The questionnaire was divided in to six sections, A to F. The information collected in each was as follows:

A Personal details: Name, sex, institution, position, years worked there, main subject taught and whether or not they have taught examination classes for 1996 and 1997;

B Use of and attitudes to feedback: Forms of feedback used (open question) and Likert (five point scale from 'agree strongly' to 'disagree strongly') items intended to measure achievement orientation, locus of control, perception of self-efficacy and attitudes to ALIS;

C About ALIS: Open questions asking how long aware of ALIS, what information received, what use made of it and the value respondents accorded to it; also stage of using ALIS ('Guttman'-type scale);

D Responsibility for students' examination performance: Pie chart to be divided according to perception of relative importance due to various factors in determining students' examination performance;

E Further comments: Opportunity for any other comments, especially comments on unclear questions;

F Consent to telephone: Whether prepared to speak further by phone.

A copy of the full questionnaire can be found in Appendix 5C (p247).

Choice of institutions

The institutions contacted were selected for one of two reasons: either because they had expressed interest in the research following a conference presentation, or because an analysis of their Physics department's examination performance showed them to be in an 'extreme' category.

The research was publicised at a conference of the Association of Principals of VIth Form Colleges (APVIC) in July 1996, at which twelve principals signed up to hear more about it. I then wrote to them with more details and received positive responses from five. These five sixth form colleges are denoted by Inst1 to Inst5 in

the remainder of this account. Two others replied to say they were not in ALIS and therefore would not be able to participate; the other seven did not reply.

The CEM centre was asked at about the same time on behalf of the Institute of Physics to identify institutions with particularly successful physics departments. Based on the average of their students' residual examination scores in physics A level over the last four years (i.e. their A level performance when the likely effect of their prior achievement is allowed for), I identified the five best and the five worst as well as five who appeared to have improved greatly and five whose performance had deteriorated. These twenty institutions were contacted by post.

After three weeks just two of them had replied, one very quickly agreeing to take part, the other declining on the grounds of lack of time. The remaining 18 institutions were randomly allocated to one of three methods of chasing up a reply: six were contacted by telephone, six were sent an additional copy of the original mailing, and for the other six, no action was taken.

Of the six followed up by phone, two agreed to participate and shortly returned the reply form, two declined (one because the school was suffering a nationally publicised internal problem and was without a headteacher, the other because they had not distributed ALIS feedback to their staff). The remaining two were still consulting the members of staff who would be involved and promised to let me know, although I heard nothing further from them.

Of those who received a written reminder, one agreed to take part and one declined, citing the restructuring process underway in their institution and pointing out that they had not yet released recent feedback to staff. I also received a positive reply from another of these some eight weeks after sending the reminder, but at that point it was too late to incorporate them into the research. No replies were received from the remaining four reminded by post or the six who received no reminder.

Of those institutions selected because of their physics results, the four who agreed to participate consisted of one sixth form consortium for a group of schools (Inst6), one F.E. college (Inst7) and two 11-18 schools (Inst8 and 9). A summary of the correspondence with each institution is shown in Table 1

Table 1: Institutions in Project 1 (with responses)

INST NO.	INSTITUTION TYPE	REASON CHOSEN	INITIAL MAILING	FOLLOW UP	DATE OF REPLY	REPLY	OUTCOME
1	SF College	APVIC	1.9.96		20.9.96	yes	in
2	SF College	APVIC	1.9.96		20.9.96	yes	withdrew 13.1.97
3	SF College	APVIC	1.9.96		20.9.96	yes	in
4	SF College	APVIC	1.9.96		13.9.96	yes	in
5	SF College	APVIC	1.9.96		13.9.96	yes	withdrew 24.2.97
6	SF Consortium	Phys ↑	14.10.96		28.10.96	yes	in
7	FE College	Phys +	14.10.96	ph 6.11	14.11.96	yes	in
8	11-18 School	Phys ↓	15.10.96	let 6.11	19.11.96	yes	in
9	11-18 School	Phys –	14.10.96	ph 6.11	21.11.96	yes	in

Key: ↑ improving; ↓ deteriorating; + high performing; – low performing; 'ph' phone; 'let' letter.

Choice of teachers

The coordinator in each institution was asked to identify all teachers of English, French, Mathematics and Physics who were currently teaching an A level examination class. The decision to use teachers of these four subjects was made for a number of reasons. Mathematics and English were the original subjects involved when ALIS began and therefore they have the largest amount of background comparison data. They are also the subjects (particularly mathematics) on which the largest amount of research has been done in the fields of school and teacher effectiveness. They are also typically the largest subject departments, which would make the administration of the project considerably easier for a given number of students. The inclusion of French and Physics was motivated partly by the desire for curriculum balance in the light of previous research on ALIS (Tymms, 1995) which found that teachers of different subjects responded differently to the feedback they got. After the identification of 'outlying' Physics departments, it was felt that their inclusion in the study would enable comparisons to be made between those at the extremes of both the 'performance' and the 'direction of change' continua. However, as can be seen from Table 1, only four of the institutions selected because of their physics department agreed to take part.

Administration and return of the questionnaires

The first batch of the revised questionnaires was sent to 108 teachers of English, French, Mathematics and Physics in six institutions between 30.10.96 and 4.11.96. Five of these institutions (all sixth form colleges) constituted the 'APVIC' sample (Inst1 to Inst5); the remaining institution (Inst6) was one of the 'Physics' sample who had sent back a very quick positive reply and was thus in time to be included with the first mailing.

The questionnaires, together with a covering letter and stamped addressed envelope with each, were sent to a coordinator for the project in each institution, who then distributed them to the appropriate teachers. After four weeks I telephoned the coordinator to let her/him know which questionnaires had not been returned by that point.

A further 48 questionnaires were sent to teachers of the same subjects in the remaining three institutions (Inst7 to Inst9) in the 'Physics' sample between 27-29th November 1996. When the request for information about teaching groups (see below) was sent (21.1.97), a note was included listing those teachers who had not yet replied.

Figure 1 shows the percentage of questionnaires returned in each institution in each of the weeks following distribution. It can be seen that the institutions differed considerably in their response rates, from a 100% return in the best (Inst4) to just 8% in the worst (Inst2 -which later withdrew from the study). In fact a chi-squared test shows that it is highly unlikely ($p = 2\times10^{-6}$) that such variation would arise by chance. One could speculate on the cause of the difference: perhaps teachers in some institutions were keener to be involved, perhaps better organised, or (more likely) the way it was presented to them and the encouragement they received to reply depended crucially on the project coordinator in the institution. It was known, however, that the institution with the lowest response rate (Inst2) had been the subject of an FEFC inspection at about the same time as the questionnaires were sent out, and this was undoubtedly a factor in the poor return. What was not known was whether similar external pressures affected any of the other institutions, but it seems likely that they may have done. Figure 2 shows that the pattern of responses for each of the subjects was broadly similar, and a chi-squared test for independence confirms this impression ($p = 0.4$).

In all, 73 questionnaires were returned (47%), but the last of these did not arrive until April! It can be seen from the graphs that a substantial proportion of the final return (21 of the 73, i.e. 29%) took more than four weeks to be received. Although the overall response rate was about what might have been expected (see, for example, Frey, 1983), the amount of time it took for the replies to come in was not anticipated, and this delay (and other delays) held up the implementation of the experiment appreciably.

Figure 1: Percentage response to initial questionnaire by institution

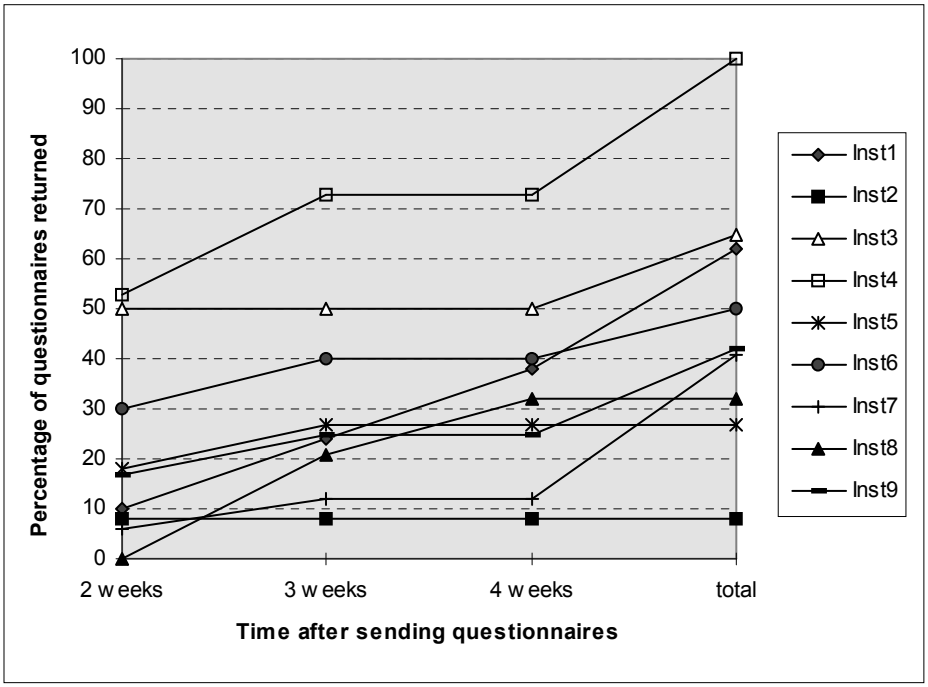
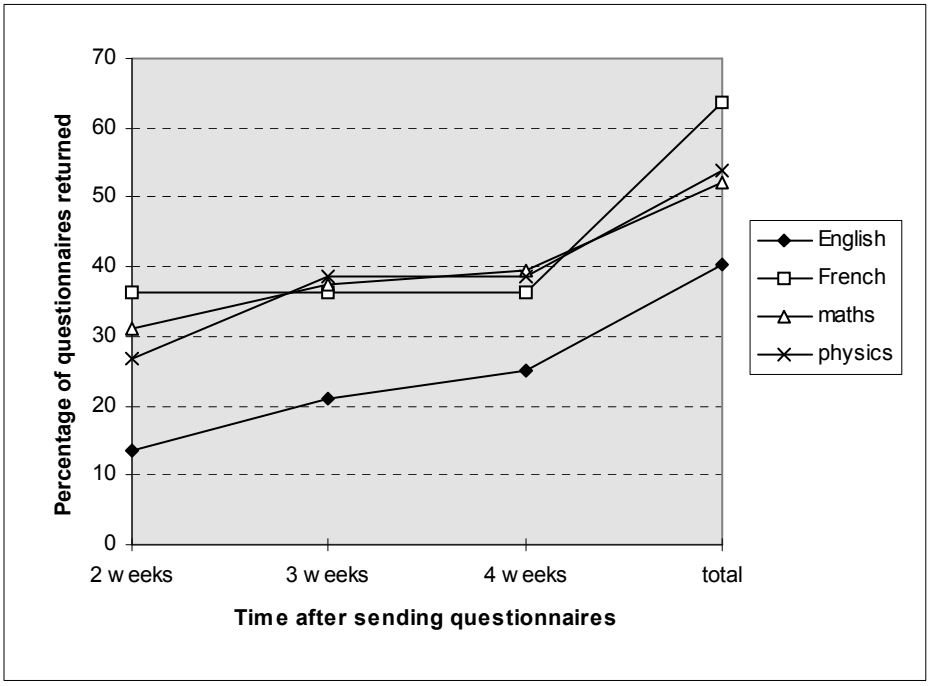


Figure 2: Percentage response to initial questionnaire by subject



5.3 IDENTIFICATION OF TEACHING GROUPS

For each of the four subjects in each institution a list of all the students in the ALIS database who had taken A level in that subject in each of the years 1994, 1995 and 1996 – and those who were due to take it in 1997 – was provided. Teachers were asked to initial beside the name of any student they had taught. These lists were sent to all nine institutions in January 1997. In response, two of them withdrew from the project: one (Inst2) immediately, citing pressure of time (partly resulting from recent FEFC inspection), the other (Inst5) after six weeks, following a series of unrelated problems with the ALIS project. These two institutions both had particularly low response rates for the initial questionnaires (2 out of 25 and 3 out of 11, respectively), so it may be that their commitment to the project was never very high. It was recognised that asking heads of department to supply this information was a substantial demand on their time.

At the beginning of March 1997, seven out of 23 eligible departments had returned the list, and the coordinator in each institution where a complete reply had not been received was contacted by telephone. In the month following the phone calls a further four returns were received, and five more arrived in the next month.

At this point in the data collection, the extent to which A level groups tended to be shared by more than one teacher became apparent. This sharing would make it difficult to identify an individual 'teacher effect', since the performance of a shared group may be more than just the sum of the individual teacher effects: there may also be an 'interaction effect' for that particular combination of teachers. In some of the syllabuses under study (e.g. modular mathematics) it might have been theoretically possible to separate the examination performance into components, each taught by a different teacher. Of course, it is arguable to what extent two separated parts of the same syllabus would really be independent – though that in itself would be an interesting empirical question. However, a large number of syllabuses would not be so separable, and as ALIS does not routinely gather modular or other component scores, collecting it would mean yet another demand on the time of the heads of department in the project and the resultant delays and attrition. For these reasons, this study did not attempt to attribute components of A level performance to individual teachers, though that would be an interesting subject for a future study.

5.4 FEEDBACK

In order to investigate the effects of giving feedback to the teachers in the project, they were randomly assigned to either the treatment group (who received feedback about the intake characteristics, attitudes and performance of the students they had taught, and 'target grades' for those about to take their examination) or the control group (who did not). Thus it was hoped that any differences between the two groups in their responses to the final questionnaire, in interviews or in the performance of their students in the subsequent exams would be attributable to the effects of the feedback.

Unit of randomisation

Because the majority of classes were taught by more than one teacher, it seemed not to be feasible to allocate individual teachers from the same department to different groups: if two teachers shared a group and one of them had the feedback and the other didn't, they would surely share the information and so 'contaminate' the control. The

decision was therefore made to allocate intact departments randomly to either treatment or control. The main disadvantage with this method was that, particularly with the relatively small numbers involved, it would be difficult to rule out institutional effects to account for any differences between the two groups. It was clear by this point that there was a potential problem with dwindling numbers, following the withdrawal of two institutions and the slow response of some of the others in providing the teaching groups information.

Assignment to treatment or control

It was decided to wait until what seemed to be the majority of departmental returns had been received and then to pair departments by subject, balancing numbers of teachers in the two groups where possible, and allocating one to the treatment group and the other to the control.

By 14th April, replies from 11 departments (with information about 30 of the teachers who had completed the initial questionnaire) had been returned. For each subject the numbers of participating teachers (shown in brackets) were as follows:

English:	Inst1 (6)	Inst6 (2)	Inst7 (2)	
French:	Inst3 (2)			
Maths:	Inst3 (7)	Inst6 (2)	Inst7 (1)	
Physics:	Inst1 (3)	Inst3 (3)	Inst6 (1)	Inst9 (1)

The departments were paired as follows:

Inst1 Physics (3)	with	Inst3 Physics (3)*
Inst6 Physics (1)*	with	Inst9 Physics (1)
Inst6 Maths (2) Inst7 Maths (1)* Inst1 English (6) Inst3 French (2)	with	Inst3 Maths (7) Inst6 English (2) Inst7 English (2)

For each pair, the department shown with a * was selected by the toss of a coin to receive the feedback. Thus 15 teachers were sent feedback and 15 were not. Shortly after this selection was made, another set of returns (from 4 departments in the same institution) was received. These were paired as follows:

Inst4 English (4)	with	Inst4 Maths (4)*
Inst4 Physics (3)*	with	Inst4 French (3)

Thus a further seven teachers received the feedback and seven did not, making a total of 22 in each of the treatment and control groups. After the second batch of feedback had been sent, one more department (Inst1 Maths) with four participating teachers returned the teaching set information, but as there was no department with which to pair them, no feedback was sent.

Content of the feedback

Teachers in the 'treatment' group received first a printout from the ALIS database listing their current students (i.e. 1997 examination entry) showing 'predicted' and target grades ('TARGETS'). The first batch of these (to institutions 1,3,6 and 7) was sent on 15.4.97; the second (to institution 4) on 29.4.97. A second dispatch containing information about students taught in the previous three years ('RESULTS', 'CLASS AVERAGES' and 'SUMMARY BY TEACHER') was sent about a week later (on 25.4.97 and 1.5.97 respectively). This feedback comprised individual student level data on value added performance in that subject as well as information about students' performance in their other subjects. Class averages for a range of intake and outcome measures (including attitudes) were provided and an overall summary calculated the average value added performance of all students taught by that teacher over the three year period. A more detailed description of the content of each part of the feedback can be found in Table 2, and a sample copy of the printouts is provided in Appendix 5D, together with the guidance notes that were included.

Table 2: Information contained in the feedback sent

PRINTOUT	SENT FOR EACH	INFORMATION CONTAINED
'TARGETS'	teaching group (1997 entry)	The average GCSE score of each student, a 'predicted' grade (with indication of likely accuracy) and 'minimum target grade' (min. grade required to gain a positive residual)
'RESULTS'	teaching group (1994-96 entry)	Average GCSE scores, A level grade, and 'value added' (residual). Also, information about each students' whole programme: number of subjects taken, overall arts/science balance, total UCAS points achieved and their value added performance averaged across all subjects. Value added in that subject and 'relative value added' (difference between subject value added and average value added for all subjects) further categorised as '+' (top 25%), '0' (middle 50%) or '-' (bottom 25%).
Graphs	Included with 'RESULTS' for each group with over three students	Two scatter graphs, one showing A level grade against average GCSE, the other showing subject value added against average for all subjects. Position of each student represented by their initials.
'CLASS AVERAGES'	teacher	Class averages of avg GCSE score, ITDA (if available), parental occupation, the percentage of female students, likelihood of staying in education, A level grade, standardised residual (value added), students' average residual in all subjects, and their attitude to the subject. An average of each of these values for all students in the department and for the whole ALIS cohort was also included.
'SUMMARY BY TEACHER'	teacher	An overall summary of the value added performance of all the students taught by that teacher over the three years.

The guidance notes tried to explain briefly how to interpret and use the data. In particular, teachers were advised to pay attention to, validate from their own perceptions and account for:

- individuals with extreme high or low value added performance
- performance of any individuals whose residual in their subject is significantly different from that in other subjects
- overall group performance and intake characteristics
- any differential effects (e.g. by ability, gender, etc.)

5.5 IMPLEMENTATION-CHECK QUESTIONNAIRE

A short questionnaire was sent shortly afterwards (15.5.97) to the teachers in the 'treatment' group in order to assess the amount of attention they had given to the

feedback and asking for comments on it (see Appendix 5E, p257). Twenty-two questionnaires were sent and teachers were reminded up to three times over the next few weeks if they had not returned theirs. Ultimately, fifteen replies (68%) were received.

5.6 FINAL QUESTIONNAIRE

This questionnaire contained some of the same Likert scale items from the initial questionnaire. Twenty-one of the original 29 items were retained because they were found to be of use in measuring the attitudes identified as significant (see Chapter 6 for the justification of which ones). To avoid confusion, these items were labelled with the letters 'A' to 'S' in order to make it clear on which questionnaire a particular response had been made. In addition, some open ended questions were included, inviting respondents to describe any changes they might have made as a result of being in the project. It also asked whether they though that class-by-class analysis should in future be sent to individual teachers, heads of department and/or the ALIS coordinator. A copy of the questionnaire can be found in Appendix 5F (p258).

Copies of this were sent to all the teachers in the 'treatment' or 'control' groups. 44 questionnaires were sent out on 12.6.97, together with a covering letter thanking them for their involvement in the project and stressing the need for all questionnaires to be returned. A copy of a recent ALIS newsletter was also sent since the inclusion of a small, unconditional gift has been shown to increase return rates (Cohen and Mannion, 1994). The questionnaire was kept to two sides of A4, since it was felt that anything longer could make some people less likely to complete it. Those who had not returned it were reminded a fortnight after dispatch, and again a week after that. Ultimately, 40 of the 44 questionnaires (91%) were returned.

5.7 FINAL INTERVIEWS

Purpose

It was intended that the interviews would achieve two things: firstly, to enable the constructs derived from the items on the questionnaires to be validated by 'triangulation': measuring the same thing with a different instrument. It was only possible to conduct a small number of interviews because of time constraints, but it was hoped that this would nevertheless provide additional evidence about the validity of the constructs used in the initial and final questionnaires. Secondly, by questioning people in a way that allowed them to describe their attitudes and perceptions in their own words and by probing in response to their answers, it was hoped that greater insight would be gained.

Sample

The sample of people to be interviewed was chosen after analysis of the results of all three questionnaires. This analysis, including the creation of the attitude constructs, is described fully in Chapter 6. It had originally been planned to choose people whose responses had placed them at an extreme on each construct and to interview them 'blind' – i.e. without knowledge of their questionnaire responses. As a substantial number of constructs had been derived from the questionnaires, it would clearly be necessary to be selective. However, the analysis of the constructs and the effects of the feedback showed that some of them appeared to be of more interest than others, and in the end it was decided to concentrate on just two: 'ease of understanding' and 'ALIS fairness'. The former was from a question on the implementation-check questionnaire which had asked respondents to say how easy to understand they had found the feedback, choosing from 'very easy', 'easy', 'moderately hard', 'hard' and 'impossible'. Although only 13 responses had been received to this question, they covered the full range, a fact which had immediately seemed to be both interesting and significant: the impact of feedback on a person who finds it 'very easy' to understand must surely be quite different from that on one who finds it 'impossible'. Moreover, when the responses to the final questionnaire were

analysed, this variable was found to be correlated with residual changes on a number of the constructs (see Chapter 6 for a full explanation of the analysis). In other words, there seemed to be significant interactions between the effects the feedback had and the ease with which it was understood. Clearly, 'ease of understanding' of the feedback was important and deserved further attention.

The other variable, 'ALIS fairness' (the extent to which individuals see ALIS feedback as fair), was chosen partly because residual change on this variable between the initial and final questionnaires was most strongly associated with 'ease of understanding' ($r = 0.52$). However, it was also of interest as the variable with the biggest treatment effect (effect size = 0.5[5]), particularly since the overall effect of the feedback seemed (if the questionnaire constructs were to be taken at face value) to be to *reduce* people's belief in ALIS as a fair indicator of performance, relative to the control group. In fact, most of the absolute change was in the control group: their belief in ALIS had apparently increased. The interpretation and implications of the changes in attitudes measured by the questionnaires are discussed more fully in Chapter 6.

One consequence of using 'ease of understanding' to select the people to be interviewed was that the sample would be limited to those in the treatment group. However, with a small number of interviews the sample could not be expected to be representative, and it seemed more important to choose people whose responses were interesting. Moreover, one of the main purposes of the whole study was to try to understand people's reactions to feedback, and clearly this would only be achieved by focusing on those who had received it. Three people at each end of the scale were therefore selected: three who had said the feedback was 'very easy' to understand, one who had said it was 'hard' and two who said 'impossible'. Unfortunately two of the three who said 'very easy' had not agreed to be contacted by phone, which opened the field to the six who had rated the feedback as 'easy' to understand (of whom one had not agreed to be contacted). At this point consideration was given to the other variable, 'ALIS fairness', and the only two whose (relative) belief in ALIS had apparently increased were chosen.

---

[5] A full explanation of how the effect size was calculated and its interpretation can be found in Chapter 6.

Others in the whole sample whose residual change in 'ALIS fairness' was large and positive were considered, and the largest one available (in the control group) was selected. Finally one 'outlier' who had stood out from an examination of the overall pattern of residual changes in the constructs was selected. This person had large residual changes (greater than two standard deviations) on eight of the nine constructs calculated, while no-one else had changed this much on more than two of them. it seemed important to know how to treat this extreme case, especially given the small size of the sample.

Timing

The interviews were conducted during September 1997. It had been hoped to complete them sooner after the feedback had been sent in order that it would be fresher in the minds of those being questioned, but the slow response of some of the returns of the final questionnaire prevented this. Clearly, any interviews conducted before the final questionnaire had been completed might have influenced the responses, and so would have made it impossible to attribute any effects to the treatment.

One consequence of this delay was that the interviewees seemed to have some difficulty answering specific questions about the feedback, and as the interviews progressed, these questions were generally omitted.

Methodology

The interviews were once again conducted by telephone, each one taking from ten to twenty minutes, and the conversation recorded and transcribed. An interview schedule was produced in order to standardise certain questions, but without restricting the interviewer's freedom to respond to comments made or to encourage respondents to talk freely. The schedule was modified slightly after two interviews when it became clear that asking respondents to place themselves on a scale of 0 to 10 in response to the questions might be to restrict them inappropriately. The first two interviews could therefore been seen as a 'pilot' study, although it is thought that the nature of most of the information gained from them is not such as to be highly

sensitive to the exact form of questioning used. A copy of the schedule used can be found in Appendix 5G (p260).

It was found to be quite hard to keep to the schedule. It seemed important to focus on what was being said and to respond to it with specific follow up questions or encouragement. This resulted in considerable deviation from a common format for all the interviews.

5.8 EXAMINATION ANALYSIS

The examination results of the students in the project departments were extracted from the database and matched with the information previously held about those students (including their teaching groups). This process was not entirely straightforward, however, since it was not unknown for a student to appear twice in the database (possibly, but not necessarily, with the same unique identifier!) and for important matching and informational variables to be different or missing. Nevertheless, this problem was solved and a program of SPSS command syntax was written for the extraction and matching of the data.

Data about students' attitudes towards the subject were also obtained from the ALIS database. The 'Attitude to Subject' scale was constructed from five Likert items on the ALIS questionnaire and scored between 1 (negative) and 5 (positive). The questionnaire was administered by ALIS in the final term of the A level course, so in many cases there would have been very little time between the teachers' receipt of the feedback and their students' completion of this questionnaire. Also, the questionnaire was not used by all institutions, since it came from the optional part of the 'Full ALIS' project. It therefore seemed that, even if providing their teachers with feedback would be expected to have an effect on students' attitudes, such an effect might well not be captured by any differences on this measure.

Missing data

There were problems with the collection of the 1997 examination results data from three of the institutions (Inst6, Inst7 and Inst9).

Inst9 had no 1997 entries in the database, either when the feedback was sent or subsequently, so no analysis could be done with their students' performance. Although the institution had been a member of the ALIS project, it withdrew while the experiment was in progress and the data could therefore not easily have been obtained. Only one teacher from that institution had been in the experiment (control group) and there were 29 student results in the years 1994-6.

A more serious loss was sustained in Inst6. Five teachers (three in the feedback group, two in the control) in three departments had taught 57 students between 1994 and 1996. The feedback on their past results, and target grades for 1997, were sent to the teachers in the feedback group, but after the 1997 examinations, their results were either missing from the database or results appeared for a completely different set of students. Clearly there was some problem with the data matching, a problem which was not resolved in time for the correct data to be included in the analysis.

In Inst7, none of the 1997 entries were in the database at the time the departments were asked to indicate which set each student was in, so no target grades were sent and no information was available about which students' teachers had received the feedback. Unfortunately, this omission was not noticed until too late, and the 1997 results were not included in the analysis. Three teachers (one in the feedback group, two in the control) in two departments were included in the experiment and a total of 25 student results from 1994-6 were involved.

There were therefore included in the analysis of the 1565 results for the years 1994-6 a total of 111 student results that came from departments whose 1997 results were not available. This represents 7% of the 1994-6 sample, a small but possibly significant proportion. When considered in terms of the number of teachers involved, the significance of the loss of data seems somewhat greater. Of the 44 teachers randomly allocated to either the feedback or control groups, information about their students' performance in 1997 was missing for 9, leaving 18 in the feedback group and 17 in the control. Unfortunately, it was discovered after allocation that a further four teachers were not teaching A level classes in 1997, so the surviving numbers of teachers with complete data both before and after the intervention were 16 in the feedback group and 15 in the control. The loss of almost 30% of the sample, in terms of the number of teachers involved, may be seen as a significant threat to making causal attributions for any differences found between the two groups.

Chapter 6

Project 1: Analysis and Interpretation of Findings

This chapter presents the results from Project 1, together with analysis, interpretation and discussion of them.

6.1 EXPLORATORY INTERVIEWS

The process of conducting these interviews has been described in Chapter 5. Transcripts of the interviews can be found in Appendix 6A (p262).

As only three interviews were conducted, and all three interviewees were heads of apparently successful departments, there is some danger of over-generalising the results. Nevertheless, they did provide evidence about the kinds of feedback people used and their attitudes towards it. The kinds of feedback mentioned in the interviews included appraisal, informal feedback from colleagues, feedback from parents, particularly via parents' evenings, from students, from OFSTED inspections, and from examination performance, whether adjusted to give 'value added' or not.

The interviewees all seemed keen to receive feedback, whether it was positive or negative, in order to evaluate their performance. For example, Peter:

> I've found both appraisals extremely useful, from the point of view of praising what I do, but also criticising some of the things I don't do, or rather criticising me for not doing certain things. You learn from that and I've taken a lot of things on board since then …

However, it did seem that most of the feedback these heads of department received was in fact quite positive. This is illustrated by Peter's comment that negative feedback from parents was 'fairly infrequent'. Positive feedback is, of course, easier to receive than negative, and it may be that their attitude towards feedback was a result of their general success and consequent tendency for the feedback to be positive. It could equally be, however, that their success was at least in part a result of their willingness to seek feedback. In response to the question asking what kinds of additional feedback he would like to get, Brian was unable to suggest anything. The other two both identified a need for more feedback from parents, particularly from those who did not tend to go to parents' evenings. However, Peter's comment that 'there is a fair amount that comes our way' suggests that he, at least, felt that his working environment was fairly rich in feedback. One interesting comment that seemed to indicate a more widespread desire for feedback was Brian's statement that all departments that he had interviewed were keen to get 'personal information' (i.e. value added analysis for each individual teacher). It is interesting also that Tim was already providing this for some of the departments in his school.

One of the main objectives of the interviews was to investigate the credibility of different forms of feedback and the extent to which teachers would perceive feedback as providing valid judgements of their performance. A number of interesting comments were made in this respect. First among them was Brian's identification of 'gut feeling' as the source of judgements about teaching quality. Although, when pressed, he was able to pick out specific features of good teaching (e.g. 'discussion amongst pupils'), he seemed reluctant to do so, believing instead in his own intuitive judgement:

> Well, I think I know … really. Some things may be pointed out, but I still think I know, and I think a lot of teachers know what is going on in their lesson. They can tell by the feel of it. The same way that I can walk in to a lesson and I think I know straight away if it's good or bad – there's an atmosphere.

For Tim, value added analysis of examination results was an important supplement to this kind of intuitive judgement:

> It often confirms what you already know, but you've got some concrete figures to back it up. That's the beauty of the value added data: not just basing things on gut feeling.

Brian, however, also commented on the danger of statistical feedback:

> … we tend to believe statistics. I suspect the people who believe it most will be those who are least happy with numbers, so we [the mathematics department] may take it a little less as gospel than some of the other departments who are less statistical.

Formal appraisal was mentioned by all three of the interviewees and was clearly an important source of feedback (see Peter's comment, above). Also mentioned was more informal feedback from colleagues. Tim referred to 'feedback I get from the members of the team' as part of the justification for describing himself as a good teacher and head of department, and Peter made a similar comment about the opinions of the members of his department. He cited particularly the need to 'keep things as open as possible' in order to encourage this feedback. Tim also stressed the informal character of his department and of the discussion that regularly occurred. Clearly, this kind of feedback will not readily be given unless it is to some extent encouraged. Peter's remark that feedback from the Head would carry more weight than other feedback may perhaps be interpreted as reflecting not so much the truth or validity of the content of the feedback as the possible consequences of the judgements being made. Feedback from the Head is important not because of his wisdom, but his power.

Feedback from parents and students was also widely mentioned. Parents' views seemed to be given more weight, and the need to seek a more representative sample than just those who came to parents' evenings was recognised. Pupils' opinions were thought to be important, but perhaps not as important as those of parents or colleagues (e.g. by Peter). Feedback from OFSTED was said to have been valuable by Brian, though comments made by him (questioning its validity) and by Tim (stressing the need to 'present the school in the best light') suggest that the importance of OFSTED may again owe more to its power than its wisdom.

Finally, a number of 'objective' measures of performance were cited as providing feedback. Tim referred to the take up of A level mathematics as a 'crude measure' of his success. Examination performance was acknowledged by Peter as, at least to some extent, an indicator of his performance. Both Brian and Tim talked about value added, but the prominence of this in the discussion may have been influenced by the fact that I initially contacted them at a conference on value added,

and by my explanations about the purpose of the interviews. Moreover, they were both the value added coordinators within their own institutions, and are unlikely to be representative of teachers in general.

6.2 INITIAL QUESTIONNAIRE

Exploratory data analysis

Recording and processing responses

For the purpose of entering and analysing the data in an SPSS file, each 'closed' question or item (i.e. those with a limited choice of outcomes) was identified as a particular variable, and each response was given a numerical code. The codings used are shown in Appendix 6B (p267). In some cases, the variable was coded in more than one way, for example, POSN ('position in the institution') was coded first on a six point scale (1 = subject teacher/lecturer; 2 = Deputy HoD/subject responsibility; 3 = Head of Department; 4 = Head (or Dep. Head) of Faculty; 5 = Senior Management; 6 = other) and then re-coded (as POSN2) for simplicity on a three point scale (1 as before as 'subject teacher', 2 and 3 combined as 'departmental responsibility', 4 and 5 combined as 'management responsibility, and 'other' now coded as 'missing'). This latter scale was expected to have more of an ordinal character, and therefore to open up the possibility of testing for associations by calculating correlation coefficients. The Likert items were also coded in different ways: initially from 1 = agree strongly to 5 = disagree strongly, and then by combining 1 and 2 as 'agree' and 4 and 5 as 'disagree' in order to test how sensitive the findings were to the (arbitrary) choice of scale on which the extent of agreement was measured.

The frequencies of each response were calculated for the nominal variables (Appendix 6C, p269) and Likert items (Appendix 6D, p271). The distribution of the sizes of each sector of the 'pie' in the question on the responsibility for students' performance was plotted on a histogram (Appendix 6E, p273). Responses to the

'open' questions – along with unsolicited written comments added at any point in the questionnaire – were transcribed and are presented in full in Appendix 6F (p276).

Characteristics of the sample

The institutions to which questionnaires were sent were a highly selected sample, having all been in the ALIS project for at least three years and having either volunteered spontaneously to take part in the research or been part of the small proportion who responded positively to a request for participants. These institutions may therefore be categorised as not only relatively experienced users of ALIS but also presumably relatively enthusiastic ones.

Of the 157 questionnaires sent out to teachers in those 9 institutions, a total of 73 were returned. However, this response rate varied considerably across institutions (see Figure 1, Chapter 5, p88), with a 100% return in one institution and only 8% in another. Owing to the generally low response rate and the way the sample was chosen, the responses cannot be taken as representative of any wider population, and any generalisations based on the questionnaires returned must be made very cautiously, if at all.

The teachers who returned the questionnaire were classified according to their main subject taught at A level (English, French, Mathematics or Physics) and their description of the position they held within the institution (subject teacher, department responsibility, management responsibility). The number of respondents in each category is shown in Table 3.

Table 3: Frequency of each combination of subject and position

SUBJECT	POSITION			
	Subject Teacher	Department Responsibility	Management Responsibility	Other/Missing
English	10	6	4	1
French	2	1	2	2
Mathematics	12	12	1	4
Physics	6	4	3	0
Missing	1	0	0	1

The pattern of responsibilities for the teachers is broadly similar in each subject: although the percentages in each category vary appreciably, the variation is no more than might be expected, given the small size of the sample. Not all of the teachers represented in this table were subsequently used in the experiment, so no further analysis of the relationship between subject and position was done at this stage.

Responses to open questions

A variety of forms of feedback were mentioned in response to question 'B', which asked for any feedback or information people had received about their job performance. Nearly three quarters (53) of the respondents named some kind of feedback, the most common being appraisal (mentioned by 26, i.e. 36%) and feedback from students (mentioned by 22, i.e. 31%). Other specific kinds of feedback mentioned were feedback from the 'line manager' (by 16), from inspectors or other observation (by 14) and from ALIS (12). It is perhaps surprising that so few (12, i.e. 17%) mentioned ALIS as a source of feedback about their job performance. All the teachers were in institutions which had been members of ALIS for at least three years, and the mention of 'ALIS' in the questionnaire title might have been expected to bias respondents towards thinking of this particular source of feedback.

The second set of open ended questions asked specifically for the information they had had from ALIS (question C2), the use they had made of it (question C3) and how valuable they had found it (question C4). 60 (83%) of the respondents mentioned some kind of information in question C2, the most common specific kinds being information about value added performance (mentioned by 28, i.e. 39%) and students' attitudes (12, i.e. 17%). All but three of those who mentioned students' attitudes also mentioned value added performance, and it should be remembered that the question did specifically prompt responses such as these by asking for information about the 'performance or attitudes of your students'. A significant number (22, i.e. 31%) referred to non-specific feedback with comments like 'the booklets' or 'all of it'. Spontaneous (i.e. not prompted by the question) mention of specific forms of feedback was therefore limited to 'students' comments', transcribed by ALIS (mentioned by 8), 'average GCSE scores' (4), 'chances graphs' (1), 'perceived learning activities chart'

(1) and the 'institution summary report' (1). Again, these figures seem quite low, given the breadth and detail of information provided by ALIS to schools. However, given one person's comment at the end of the questionnaire (section E) that, 'It may have been helpful to have been reminded of the variety of analyses which ALIS provides', and the likely amount of time and thought spared from the busy teaching day for a questionnaire such as this, the failure to mention specific types of feedback may not necessarily mean that people had not received – or even valued – them.

Just under half the respondents (33, i.e. 46%) were able to describe some way they had used the information received from ALIS (question C3), the majority of these (20) mentioning some analysis of the examination performance of their students. Of these, five specifically mentioned using a 'set by set' analysis. Other uses included target setting, identifying under-achievers, and analysing attitudes and comments.

The question (C4) asking how valuable the feedback was produced a range of responses, most of which did not answer the specific question in its strict sense. The answers of 14 people could be taken as saying that at least part of the feedback was of some value, but only three of these endorsed it without reservations, the others typically describing it as of 'moderate value' or finding value in only specific parts of it. On the other hand, 20 people attributed little or no value to it, typically replying, 'not very', 'little' or 'not at all'. A further seven responses used the word 'useful' to describe the feedback and 10 used the word 'interesting'. A number of specific criticisms were made of the ALIS feedback, including questioning the use of GCSE average scores to predict A level grades, the perception that the feedback generally tells you what you know already and the model's perceived oversimplification of the complex issue of accounting for student performance. Some of these reservations were also raised in responses to the final 'any additional comments' question (section E). The overall impression gained about the value these teachers attributed to the feedback they received from ALIS is that while some found it interesting, useful or valuable in part, many did not, and there were some fairly strongly held reservations about its use; very few were prepared to endorse it unconditionally.

Finally, respondents were asked to make 'any additional comments' or to comment on the questionnaire itself (section E). As well as the issues already mentioned above, a number of significant comments were made here, generally critical of either specific questions in the questionnaire (see next section, below) or of

specific aspects of ALIS, often also suggesting improvements. Although the general level of comments suggested a fairly high degree of familiarity with ALIS, four of the responses were to the effect that the person did not know enough about it to have been able to complete the questionnaire adequately.

Interpretation of attitudes towards ALIS: inter-rater consistency

The written comments from all parts of the questionnaire were then classified according to whether they suggested a person whose perception of ALIS feedback was 'generally positive', 'generally negative' or 'mixed/neutral/not clear'. This was done in order to get an overall picture of people's attitudes towards ALIS from their open-ended written comments and to see the extent of agreement between attitudes inferred from written comments and the attitude scales derived from the Likert-type items (see below). However, since the classification of questionnaires into 'positive', 'negative' or 'neither' was inevitably to some extent a subjective one, the classification was made first by me and then independently by three other teachers, none of whom were involved in the study. Analysis of these ratings provided some interesting findings.

Each of the four people was provided with a transcript of the comments made by the questionnaire respondents in response to questions C2, C3, C4 and E. They were given the criteria shown in Figure 3 and asked to classify each questionnaire into one of the three categories.

Figure 3: Criteria for classifying attitudes to ALIS

'**generally positive**'	If they have described the information as valuable or interesting or have made considerable use of it;
'**generally negative**'	If they have described it as not valuable or have made substantial criticisms of it;
'**neutral/mixed**'	If their attitude is neutral, not clear or a mixture of positive and negative, or if they have not received enough information from ALIS to comment;

The number of questionnaires for which each pair of raters agreed is shown in Table 4. Of the 72 questionnaires rated, no pair agreed on the classification of more than 51 and the mean number of agreements was 43.5, or just over 60%. All four raters agreed on the classification of 23 of the questionnaires (32%). Of these unanimously agreed questionnaires, eight were classified as 'positive', five 'neutral' and ten 'negative' in their attitudes towards ALIS. Two questionnaires had attracted opposite ratings from at least one pair of raters, and the remainder had either a mixture of positive and neutral or a mixture of negative and neutral ratings. Finally, one of the raters (number 2) was asked to rate the questionnaires again, about a fortnight later, producing 54 agreements with her 72 previous judgements.

Table 4: Number of agreements between raters on attitudes towards ALIS

RATER:	2	3	4
1	40	51	41
2		46	44
3			39

Total no. of questionnaires = 72.

It is hard to resist the conclusion from these data that the interpretation of open comments on the questionnaire is somewhat problematic. Even an apparently simple question about respondents' attitudes towards ALIS, 'Were they generally positive or negative?', cannot really be answered reliably from what they wrote on the questionnaire. For about a third of the questionnaires, all the raters would have agreed on the answer to this question; for the other two thirds, however, the answer would depend on who you asked to interpret the comments made. It should be said, though, that almost all of the disagreement was about whether to classify a response as neutral or not; only two questionnaires were rated as positive by one person and negative by another.

It may be that a simple positive/negative classification was not really appropriate for the complex views that people had about ALIS. Many of those who were positive about some aspect of it were less enthusiastic about some other, so it is perhaps not surprising that they could not all be neatly categorised. Equally, it would have been

difficult to draw up a wholly unambiguous set of criteria for classifying the responses, but it certainly arguable that the criteria given could have been improved in this respect. Nevertheless, this analysis of inter-rater consistency does illustrate some of the difficulties of interpreting open comments. Many of the interpretations given above should be treated with appropriate caution.

These ratings of attitude towards ALIS were also used to create an average score for each questionnaire, and these scores were compared with the attitude constructs derived from the Likert-type items. This process of 'triangulation' is described on page 129.

Questions which were perceived as problematic

Some of the questionnaire items appeared to have been seen as problematic, either because of comments made about the item, or because of a high rate of non-response (or both). This could indicate that the item was seen as ambiguous or inappropriate in some way, and it seems likely that even those who answered the question and did not comment may have shared some of this feeling about the item. The interpretation of these items must therefore be treated with some caution.

The following items either had more than two missing responses or received at least one comment which suggested they were problematic:

- B04 ('I believe I am a good teacher') (2 non-responses, 2 comments). Comments suggest some concern with ambiguity. However, ambiguity need not prevent a statement from measuring attitudes satisfactorily (Oppenheim, 1992, ch10)
- B10 ('The ALIS data on attitudes do not tell us anything worthwhile') (4 non-responses, 1 comment)
- B11 ('I prefer tasks in which I can see how well I am doing') (0 non-responses, 1 comment)
- B19 ('My institution gets very little benefit from being in ALIS') (3 non-responses, 3 comments). Comments made (and later comments in section C) suggest these respondents simply did not know the answer.
- B23 ('There are too many errors in the feedback provided by ALIS for their findings to be reliable') (4 non-responses, 3 comments). Again,

comments suggest this item was omitted at least partly from lack of knowledge about it.

- B27 ('Doing well is more important to me when I am being assessed') (2 non-responses, 1 comment)
- B29 ('I think the Head/Principal should not use ALIS results in staff appraisal') (3 non-responses, 3 comments). The double negative created by disagreeing with this item may have been confusing.

Thus it seems that in many cases of non-response, people excused themselves on the grounds that they did not know enough to be able to answer properly. This was particularly the case when the item related to some aspect of ALIS with which they were not familiar.

Items with low discrimination

Some of the questionnaire items received only a very limited range of responses or provoked the same response from a large majority of respondents. There are two possible explanations for this. It could be that the item failed to discriminate between respondents who were actually different with respect to the construct that was intended to be measured. In this case the underlying construct may be seen as appropriate, but its measurement, by the item in question, inadequate. Examples of this would be items whose meanings were so unclear that it would be hard either to agree or disagree, or which were worded such that almost everyone would agree with the statement. Alternatively, it might be that, for the particular sample used, the construct itself would not discriminate among the respondents, even if it were well measured. In this case the lack of range in responses could reasonably be interpreted as homogeneity of the sample: everyone gave the same response because they were all essentially the same with respect to the construct in question. Of course, it is possible that lack of discrimination could result from a combination of measurement inadequacy and sample homogeneity.

In examining the frequencies of responses, it was noticed that two of the items (B04 and B08) had attracted responses on only three of the five possible choices: no-one had disagreed (or disagreed strongly) with these statements. However, it was also

noticed that some of the other items that had attracted responses on four or even all five of the possible choices had nevertheless had only one or two people choosing the less popular values. It seemed sensible to include these items among those with 'low discrimination' in order not to give undue weight to the responses of one or two individuals. Eight of the 29 items had attracted over 95% of the responses to just three of their five choices. In all of these bar one (B01), over half the respondents had chosen the same single response. For this item and for one other (B11), over 85% of the respondents had chosen one of just two responses. These nine items, together with the frequencies of each response, are shown in Table 5, ordered by the percentage of people who chose one of the top three responses for each.

Table 5: Frequencies of responses to items with low discrimination

ITEM	FREQUENCY					
	agree strongly				disagree strongly	
	1	2	3	4	5	tot
B08. Receiving feedback can help me to improve what I am doing.	14	50	7	0	0	71
B04. I believe I am a good teacher.	15	45	10	0	0	70
B16. If the students I teach perform badly, it is their fault.	1	10	39	20	0	70
B09. If a student who does not much like my subject joins my class, I can usually help him or her to enjoy it more.	4	29	37	2	0	72
B01. I like to receive objective feedback about the quality of my work.	32	29	9	2	0	72
B25. I feel confident about the quality of my work.	12	47	10	2	0	71
B07. My effectiveness as a teacher depends on how I choose to teach.	13	39	15	3	0	70
B23. There are too many errors in the feedback provided by ALIS for their findings to be reliable.	1	19	35	11	2	68
B11. I prefer tasks in which I can see how well I am doing.	4	35	29	3	1	72

Correlations among items

The correlation coefficients (Pearson product moment) between all pairs of items which could arguably be classified as measured on an ordinal (or better) scale – or were binary variables – were calculated, and are shown in Appendix 6H (p291). It is accepted that the strict conditions for the use of these coefficients may well not be met by some (or indeed all) of these variables, but it was a broad indication of the extent to which any pair of variables interacted which was sought, rather than a

precise absolute measure of correlation or a judgement about the level of statistical significance achieved. So-called 'non-parametric' measures of correlation were also calculated (Spearman's and Kendall's coefficients) for the same variables, and correlations were also recalculated for the Likert items recoded on a three point scale. A broad measure of agreement was found from all methods, particularly with respect to the relative sizes of correlation coefficients.

Items with few associations

For each variable, the number of correlations with absolute value above 0.3, the number above 0.5 and the number above 0.7 was found. These thresholds were arbitrary, though it was felt that a correlation below 0.3 indicated that there was no relationship of any significance between the two variables. Correlations above 0.3 begin to indicate an association; those above 0.5 show a moderate relationship; those above 0.7 show a fairly strong association. All correlations above 0.3 were statistically significantly different from zero with $p < 0.05$ (and $p < 0.01$ in almost all). However, since the conditions for using a test of this kind were not met, the level of statistical significance cannot be taken as precise. Table 6 shows the number of correlations in each category for each questionnaire item.

Table 6: Number of substantial correlations for each variable

VARIABLE	NUMBER OF CORRELATIONS		
	.3 < r < .5	.5 < r < .7	r > .7
SEX	●		
YRS	●		
B01	●●●●	●	
B02	●●		
B03		●●●	
B04	●●		●
B05	●●●		
B06	●●●●●●●●●	●	
B07	●		
B08	●●●●		
B09	●●●●		
B10	●●	●	
B11	●●●●		
B12	●●●●●	●●	

B13	●●●●●●●		
B14	●●●●●●	●●	●
B15	●●●●●	●	
B16	●●		
B17	●●●	●●●	
B18	●●		
B19	●●●●●●	●	
B20	●●●●●●	●	●
B21	●●	●●●●	
B22	●●●●●	●●	
B23	●●		
B24	●●●		
B25	●●●●		●
B26	●●●●●●	●●	
B27	●		
B28	●		
B29	●●●●●●		
AWARE	●●		
STAGE	●●●		
RESP_ABL	○○○○	○	
RESP_BGD		○	
RESP_CHR	●●○		
RESP_TCH	○		
RESP_SCH	○		
RESP_OTH	○		

● *= 1 correlation;* ○ *= correlation (partly) arising from constraints on responses.*

Items which had only a small number of these 'significant' correlations were either not measuring anything consistently (i.e. their responses were effectively at random), or they were not measuring anything that was being measured by the other items in the questionnaire, or they were failing to discriminate adequately (see above). The following items warranted extra examination on the grounds of having few significant correlations:

- SEX (male/female) and YRS ('time in that institution'). These are both reporting simple facts about the respondents and may thus be believed to have relatively high validity and reliability (but see Belson, 1981, for evidence that the interpretation of even 'simple' factual questions is extremely problematic). The lack of correlation with other items suggests that these variables are not strongly associated with any of the attitudes measured.
- Three of the Likert scale items, B07 ('My effectiveness as a teacher depends on how I choose to teach'), B27 ('Doing well is more

important to me when I am being assessed'), and B28 ('When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough'), had only one correlation above 0.3. A further four, B02 ('I am always keen to have my performance assessed'), B16 ('If the students I teach perform badly, it is their fault'), B18 ('I usually seek information with which to judge whether I am achieving what I want to'), and B23 ('If the analysis by ALIS shows that a particular department has a high score, then I will believe that there has been some good teaching in that department'), had just two correlations greater than 0.3 each. Since the content of these statements appears to be similar to that of the other Likert scale items, it may well be that their meanings were not clear to the respondents, or that something about the way they were worded caused respondents to be influenced by some feature other than what was intended. Either way, it is hard to interpret responses to these items with confidence, and it may be safer to remove them from further analysis.

- AWARE ('time aware of ALIS'). This item may have failed to discriminate sufficiently between respondents, since a large majority (45) said they became aware of ALIS 'more than three years ago', and almost all (25) of the remainder (27) ticked 'between one and three years ago'. Thus, one would have expected low correlations to be found, even if the amount of time a person had been aware of ALIS was actually significantly related to other questionnaire items. However, it could equally be that 'time aware of ALIS' was simply not associated with other characteristics measured in the questionnaire. Either way, this item seemed to have measured the underlying construct only very crudely, and may therefore be of limited use in further analysis.
- STAGE ('Stage of using ALIS'). This item was an attempt to measure the extent to which teachers had assimilated the ideas behind the ALIS project into their thinking and incorporated its use into their practice (see Chapter 5). A relatively high rate (8.3%) of non-response to this question suggests that the categories offered may not have been

perceived as appropriate. It may be that the stage of development of a person's use of ALIS was not related to their perceptions of and attitudes towards ALIS and feedback in general. However, evidence from the final questionnaire (see below) suggests that this was not the case, and the more likely explanation for the scarcity of correlations between this item and others in the questionnaire is that the question failed to measure this attribute adequately. It may be that with more extensive and detailed piloting - possibly involving the construction of a proper Guttman scale (McKennell, 1977) - the question could have been developed in order to better measure what appears to have been a significant factor.

- The 'pie-chart' items, RESP_ABL to RESP_OTH (amount of responsibility for students' performance attributed to 'ability', 'home background', 'character', 'teacher', 'school' or 'other'). Almost all the significant correlations with these variables were among each other, and therefore (at least to some extent) spurious, since the sizes of the sectors were not independent (the sum of the proportions of responsibility attributed to each factor was constrained to be 100%). Once again, it is not clear whether the underlying construct was irrelevant, or whether it was simply poorly measured. Either way, the interpretation of the measured variable is problematic.

Construction of attitude scales

It is well established (e.g. McKennell, 1977; Oppenheim, 1992) that attitudes cannot reliably be inferred from responses to a single item or question. Hence, if attitudes were to be successfully measured by the questionnaire, it would be by combining items to form an attitude scale. Cronbach's alpha (Cronbach, 1951) provides an indication of the internal consistency of such a collection of items, i.e. the extent to which responses to them can be predicted from the responses to other items in the scale. It is thus clearly desirable to maximise the alpha value for a scale, and items may be added to or removed from the scale in order to achieve this.

However, the value of alpha calculated from the questionnaire responses may be thought of as only an estimate of the 'true' population value, and thus subject to a sampling error. In other words, if the same questionnaire were returned by a different sample from the same population, the value of alpha obtained would be expected to vary somewhat. The amount of variation to be expected may be expressed in terms of a confidence interval. Since I was unable to find a known formula for calculating a confidence interval for Cronbach's alpha, I used the non-parametric technique of bootstrapping (Efron and Tibshirani, 1993) (see Annex: Confidence intervals for Alpha). The implication of treating each calculated value of alpha as a parameter estimate in this way is that a small change in alpha brought about by changing the makeup of an attitude scale might well not be reproduced with a different sample. However, the bootstrapping was unfortunately not done until after the initial analysis of the questionnaire, and the decisions had already been made about which items to include in the final (post-test) version. It is likely that had I done the bootstrapping first, and so been more conscious of the confidence interval associated with each calculated alpha value, I might well have been less willing to remove an item from a scale for the sake of a small increase in alpha if it seemed otherwise to be appropriately included.

Since the purpose of measuring these attitudes was to investigate relationships between them and individuals' responses to feedback, it was felt that a reliability (alpha) of 0.7 or above would be sufficient. However, it should be noted that Cronbach's alpha provides no guarantee of the stability of an apparent 'attitude' over time and in different contexts (i.e. *test-retest reliability*), nor does an 'attitude' measured in this way necessarily equate with other manifestations of what might be thought to be the same 'attitude' (i.e. its *concurrent validity*).

It is a necessary condition of establishing validity that the attitude scale must be interpretable. If it is not clear what characteristic the scale is measuring, then any subsequently demonstrated similarity between that scale and the 'same' characteristic measured in some other way can really only be used to provide a *post hoc* interpretation of the scale, rather than a true 'triangulation'. In grouping items together to form an attitude scale, it is therefore important that they should have meaning in common. Even if the inclusion of an item leads to an increase in alpha, it

should not be added unless the augmented scale thereby created remains readily interpretable.

Two criteria were therefore used for the acceptability of attitude scales: *consistency* (as measured by Cronbach's alpha) and, more subjectively, *interpretability* (face validity). With these criteria, a number of different ways of combining items into attitude scales were used:

Face validity

A review of the feedback literature (see Chapter 3) had identified certain attributes which had been found to be significant in mediating the effects of feedback on performance. These included individuals' perceptions of self-efficacy, their achievement orientation and locus of control. In addition, it was felt that teachers' attitudes towards ALIS might also affect their responses to the feedback they received. The Likert scale items (B01 to B29) were originally included in the questionnaire in order to measure these four attributes.

Thus it seemed to be a sensible starting point for constructing attitude scales to group together the items which had been intended to capture the same characteristic, and to investigate the consistency of the scale which would be produced by combining them in this way. For each of the four intended attributes, a list of items which appeared from their meaning to be measuring that quality (i.e. on the basis of *face validity*) was drawn up. These lists are shown in Table 7 with items that were seen as central to that attribute shown in bold, and those which were felt to be more loosely connected also listed. If an item was expected to be correlated negatively with the others in that scale, then that item was inverse coded (i.e. agree strongly = 5, disagree strongly =1) in order that the scale could be constructed simply as the sum of the codings of the component items. The value of Cronbach's alpha was calculated with all the items included, and again with each item omitted in turn. If the removal of an item led to a higher alpha, it was removed and the process repeated until no further removal increased the value of alpha. Items removed in this way are shown preceded by a 'x', and the order of their exclusion and corresponding alpha values shown. The mean of the codings of all the items retained in each attitude scale was calculated and given the variable name shown (Table 7).

Table 7: Attitude scales based on face validity of items

SELF-EFFICACY: *the extent to which individuals perceive themselves as effective teachers*.

Items:

xB3. The exam results of the students I teach reflect my ability as a teacher.
B4. I believe I am a good teacher.
B14. I am worried that feedback about my teaching performance could be used against me. (Inverse coded)
B15. I often have doubts about whether I am doing a good job. (Inverse coded)
B17. The quality of my teaching is reflected in the exam success of my students.
B20. I am concerned that information from ALIS could be used to check up on me. (Inverse coded)
B25. I feel confident about the quality of my work.
B26. If ALIS gave me information about my teaching performance I would find it quite threatening. (Inverse coded)

All items:	$\alpha = 0.77$	(8 items)
Remove B03:	$\alpha = 0.78$	(7 items: mean = SELF_EFF)

ACHIEVEMENT ORIENTATION: *the extent to which individuals attach importance to their performance and value performance feedback.*

Items:

B1. I like to receive objective feedback about the quality of my work.
B2. I am always keen to have my performance assessed.
B5. I do not like situations in which I am being judged. (Inverse coded)
B6. If ALIS gave me information about my teaching performance I would find it useful and informative.
xB8. Receiving feedback can help me to improve what I am doing.
xB11. I prefer tasks in which I can see how well I am doing.
B14. I am worried that feedback about my teaching performance could be used against me. (Inverse coded)
xB18. I usually seek information with which to judge whether I am achieving what I want to.
B21. I feel anxious when I am evaluated. (Inverse coded)
B26. If ALIS gave me information about my teaching performance I would find it quite threatening. (Inverse coded)
xB27. Doing well is more important to me when I am being assessed.
B29. I think the Head/Principal should not use ALIS results in staff appraisal. (Inverse coded)

All items:	$\alpha = 0.73$	(12 items)
Remove B11:	$\alpha = 0.75$	
Remove B27:	$\alpha = 0.76$	
Remove B18:	$\alpha = 0.77$	
Remove B8:	$\alpha = 0.79$	(8 items: mean = ACH_ORN)

LOCUS OF CONTROL: *the extent to which individuals perceive success or failure as within their control.*

Items:

B3. The exam results of the students I teach reflect my ability as a teacher.
xB7. My effectiveness as a teacher depends on how I choose to teach.
xB9. If a student who does not much like my subject joins my class, I can usually help him or her to enjoy it more.
B12. I am responsible for the exam performance of my students.
xB16. If the students I teach perform badly, it is their fault. (Inverse coded)
B17. The quality of my teaching is reflected in the exam success of my students.
B22. The A level grades that students get depend on who teaches them.
xB28. When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough.

All items:	$\alpha = 0.69$	(8 items)
Remove B09:	$\alpha = 0.71$	
Remove B07:	$\alpha = 0.74$	
Remove B28:	$\alpha = 0.77$	
Remove B16:	$\alpha = 0.83$	(4 items: mean = LOC_CTRL)

ATTITUDE TO ALIS: *the extent to which individuals perceive ALIS feedback as valid and worthwhile.*

Items:

B6. If ALIS gave me information about my teaching performance I would find it useful and informative.
B10. The ALIS data on attitudes do not tell us anything worthwhile. (Inverse coded)
B13. The value-added scores (residuals) calculated by ALIS are a fair way of measuring how well students have done.
B19. My institution gets very little benefit from being in ALIS. (Inverse coded)
B20. I am concerned that information from ALIS could be used to check up on me. (Inverse coded)
xB23. There are too many errors in the feedback provided by ALIS for their findings to be reliable. (Inverse coded)
xB24. If the analysis by ALIS shows that a particular department has a high score, then I will believe that there has been some good teaching in that department.
B26. If ALIS gave me information about my teaching performance I would find it quite threatening. (Inverse coded)

All items:	$\alpha = 0.75$	(8 items)
Remove B23:	$\alpha = 0.76$	
Remove B24:	$\alpha = 0.77$	(6 items: mean = ATT2ALIS)

KEY: ***Items in bold type*** *- believed a priori to be strongly related to attitude construct.*
Items not in bold - believed a priori to be loosely related to attitude construct.
x before item - removed to increase alpha reliability of scale.

It can be seen that all four 'attitudes' can be satisfactorily measured by the items which were intended to measure them with alpha values around 0.7 or better, and that by excluding a small number of items from each scale that alpha value can be increased in each case to around 0.8.

Factor analysis

An alternative way of grouping the items was to use factor analysis (Norusis, 1985). This provides a method of identifying underlying constructs and thereby summarising a large number of variables with a smaller number of factors.

Initially all the ordinal variables (Likert items, AWARE, STAGE and the pie chart items) were included, but the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO) was found to be only 0.24. This measure indicates the extent to which correlations between pairs of variables can be explained by the other variables, and values below 0.5 are unacceptable for factor analysis (Kaiser, 1974). Measures of sampling adequacy for each variable (MSA_i) were also calculated to indicate the contribution each made to the KMO measure (Norusis, 1985). Variables with the lowest MSA_i were progressively dropped until the KMO measure reached 0.5 (which is nevertheless described by Kaiser (1974) as 'miserable'), then 0.6 ('mediocre') and 0.7 ('middling'). It was not found to be possible to raise the KMO measure above 0.79 to the 0.80 threshold of 'meritorious' by any further removal of items. Also, it was found that the choice of which variables to include at the beginning made a significant difference to the order in which variables were dropped, so it may be that a different starting point could have resulted in a higher eventual KMO measure.

However, when the factor analysis was done, the factors produced seemed to be fairly constant (in terms of the relative sizes of the factor loadings of each of the items on each factor) whatever the starting point, KMO value, method of factor extraction or rotation of factors. In particular, factors were extracted using either Principal Components Analysis or Alpha Factoring (which maximises the alpha reliability of the factors) and rotated orthogonally using VARIMAX and obliquely using OBLIMIN (Norusis, 1985). Similar results were found in all cases.

Table 8 shows a typical example of the results of factor analysis. The variables used were the Likert items (with B07, B11, B16, B23, B24 omitted) and STAGE, which produced a KMO value of 0.60. Seven factors were extracted by Principal Components Analysis, which accounted for 70% of the variance of these variables. The factors were rotated using the OBLIMIN algorithm. For each factor the items which had (absolute) factor loadings of 0.3 or greater are shown, in decreasing order of their loading. The alpha reliability of the scale constructed by simply adding the

scores for all the items (inverse coded if the loading was negative) was calculated. Once again, items were removed from the scale (shown preceded by 'x') if doing so increased the value of alpha for those remaining (Table 8).

Table 8: Grouping of items by factor analysis

FACTOR 1: (21.9% of variance)

Items:

B14. I am worried that feedback about my teaching performance could be used against me (0.86)
B20. I am concerned that information from ALIS could be used to check up on me (0.81)
B21. I feel anxious when I am evaluated (0.68)
B26. If ALIS gave me information about my teaching performance I would find it quite threatening (0.51)
xB18. I usually seek information with which to judge whether I am achieving what I want to (0.40)
xB17. The quality of my teaching is reflected in the exam success of my students (-0.36)
xB29. I think the Head/Principal should not use ALIS results in staff appraisal (0.35)
xB15. I often have doubts about whether I am doing a good job (0.34)
xB12. I am responsible for the exam performance of my students (0.33)

All items:	$\alpha = 0.70$	(9 items)
Remove B12:	$\alpha = 0.74$	
Remove B18:	$\alpha = 0.80$	
Remove B29:	$\alpha = 0.80$	
Remove B17:	$\alpha = 0.83$	
Remove B15:	$\alpha = 0.85$	(4 items)

FACTOR 2: (13.4% of variance)

Items:

B22. The A level grades that students get depend on who teaches them (0.81)
B17. The quality of my teaching is reflected in the exam success of my students (0.76)
B3. The exam results of the students I teach reflect my ability as a teacher (0.76)
B12. I am responsible for the exam performance of my students (0.71)
xB8. Receiving feedback can help me to improve what I am doing (0.42)
xB5. I do not like situations in which I am being judged (0.39)

All items:	$\alpha = 0.74$	(6 items)
Remove B5:	$\alpha = 0.80$	
Remove B8:	$\alpha = 0.83$	(4 items)

FACTOR 3: (9.3% of variance)

Items:

B25. I feel confident about the quality of my work (0.87)
B4. I believe I am a good teacher (0.72)
xB5. I do not like situations in which I am being judged (0.63)
xB18. I usually seek information with which to judge whether I am achieving what I want to (0.47)
xB15. I often have doubts about whether I am doing a good job (-0.39)

All items:	$\alpha = 0.62$	(5 items)
Remove B18:	$\alpha = 0.65$	
Remove B5:	$\alpha = 0.69$	
Remove B15:	$\alpha = 0.84$	(2 items)

FACTOR 4: (7.8% of variance)

Items:

B1. I like to receive objective feedback about the quality of my work (0.72)
B2. I am always keen to have my performance assessed (0.72)
B26. If ALIS gave me information about my teaching performance I would find it quite threatening (-0.65)
B6. If ALIS gave me information about my teaching performance I would find it useful and informative (0.48)
xB8. Receiving feedback can help me to improve what I am doing (0.45)
xB18. I usually seek information with which to judge whether I am achieving what I want to (0.34)

All items:	$\alpha = 0.70$	(6 items)
Remove B18:	$\alpha = 0.72$	
Remove B8:	$\alpha = 0.74$	(4 items)

FACTOR 5: (6.6% of variance)

Items:

B10. The ALIS data on attitudes do not tell us anything worthwhile (0.83)
xB9. If a student who does not much like my subject joins my class, I can usually help him or her to enjoy it more (-0.77)
B19. My institution gets very little benefit from being in ALIS (0.61)
xB28. When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough (0.34)

All items:	$\alpha = 0.52$	(4 items)
Remove B28:	$\alpha = 0.76$	
Remove B9:	$\alpha = 0.79$	(2 items)

FACTOR 6: (5.9% of variance)
Items:
xB28. When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough (-0.68)
B13. The value-added scores (residuals) calculated by ALIS are a fair way of measuring how well students have done (-0.65)
B29. I think the Head/Principal should not use ALIS results in staff appraisal (0.56)
xB15. I often have doubts about whether I am doing a good job (0.55)
B6. If ALIS gave me information about my teaching performance I would find it useful and informative (-0.42)
xB21. I feel anxious when I am evaluated (-0.30)

All items:	$\alpha = 0.56$	(6 items)
Remove B21:	$\alpha = 0.61$	
Remove B15:	$\alpha = 0.66$	
Remove B28:	$\alpha = 0.69$	(3 items)

FACTOR 7: (5.2% of variance)
Items:
B27. Doing well is more important to me when I am being assessed (-0.73)
xSTAGE (stage of using ALIS) (0.67)
xB18. I usually seek information with which to judge whether I am achieving what I want to (-0.55)
B19. My institution gets very little benefit from being in ALIS (0.53)

All items:	$\alpha = 0.51$	(4 items)
Remove STAGE:	$\alpha = 0.52$	
Remove B18:	$\alpha = 0.55$	(2 items)

Note: Factor analysis and calculation of Cronbach's alpha based on n = 72 responses, less a small number of missing responses on some items.

It can be seen that the items which were rejected in order to increase the consistency of the scale formed from each factor were generally also those which had the lowest loadings on that factor. After these items had been removed, an attempt was made to interpret the scale produced, and each variable so formed was given a name.

Table 9: Interpretation of constructs from factor analysis

FACTOR	NAME	INTERPRETATION
Factor 1:	Feedback-Anxiety	The extent to which individuals are anxious about receiving feedback.
Factor 2:	Responsibility	The extent to which individuals take responsibility for their students' performance.
Factor 3:	Self-Confidence	The extent to which individuals feel confident about their effectiveness as teachers.
Factor 4:	Feedback-Desire	The extent to which individuals desire performance feedback.
Factor 5:	ALIS-Value	The extent to which individuals see ALIS as of value.[6]
Factor 6:	ALIS-Fairness	The extent to which individuals see ALIS feedback as fair.

Note: Factor 7 not only had rather low internal consistency ($\alpha = 0.55$), but did not seem to be readily interpretable, and was therefore omitted.

Cluster analysis

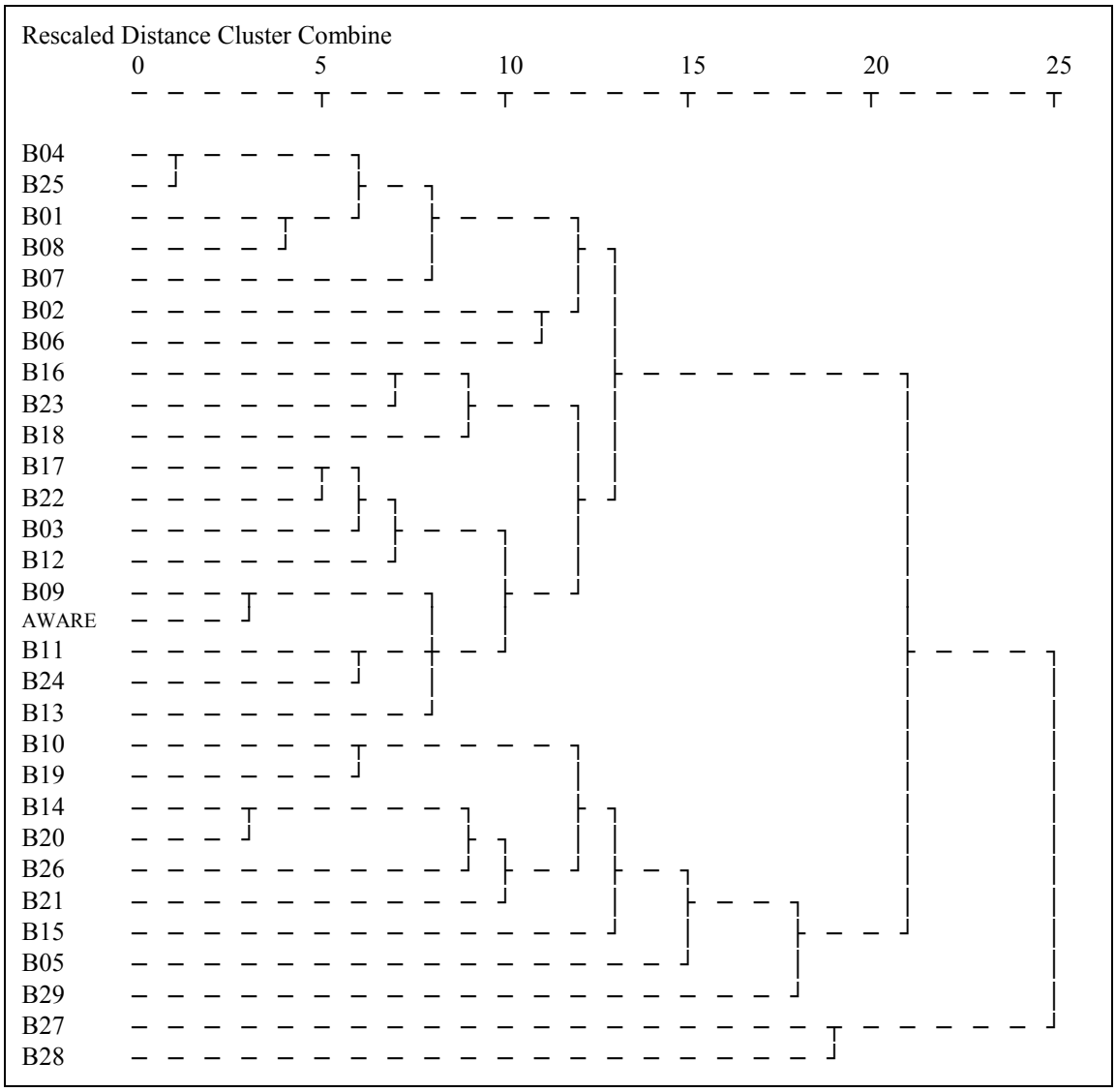
Hierarchical Cluster Analysis (Norusis, 1985) provides yet another way of classifying the variables into clusters based on the correlations among them, and can be used as a basis for constructing attitude scales (McKennell, 1977). This procedure was performed with the same starting set of variables as were used in factor analysis. Once again, the pie-chart variables failed to cluster with any of the others, although in this case AWARE (time aware of ALIS) did cluster quickly, while STAGE (stage of using ALIS) did not. The latter variable and all the pie-chart variables were therefore dropped from the analysis.

The dendrogram (Norusis, 1985) showing the rescaled distance at which the clusters combined is shown in Figure 4. Distances were based on average linkage between groups, using squared Euclidean measure. It can be seen that although there are some elements of strong clustering, the overall pattern is for gradual accumulation, rather than forming distinct clusters. Also, when the internal consistency (Cronbach's alpha) of each cluster was calculated, the values were found to be generally quite low. In fact most values of alpha were below 0.5 and values above 0.6 seemed to be limited

[6] Note that as the factors originally came out of the factor analysis, this factor should have the opposite meaning. However, for the sake of consistency, it seemed more sensible to reverse the coding of its constituent items and interpret the construct as the extent to which individuals see value in ALIS, rather than the extent to which they fail to do so.

to clusters which comprised items which had previously been associated in the attitude scales derived from factor analysis (see p120). It may therefore be said that cluster analysis failed to provide any new insights into how the items might be grouped to produce satisfactory attitude scales.

Figure 4: Dendrogram showing distances between clusters of items



'Likert' approach

Finally, one further method of constructing attitude scales discussed by McKennell (1977) was used. The Likert method of scale construction selects items which have the highest correlations with the scale total, and thus maximises the

average of the item-total correlations. This contrasts with the 'alpha' technique, used above, in which the average of the individual item-item correlations (and thereby the alpha coefficient) is maximised. McKennell (1977) argues that the alpha approach is generally to be preferred. However, after putative scales had been constructed using face validity and factor analysis, and then modified to maximise alpha, the Likert method was used to check that no other item – previously overlooked or excluded – could be combined with any attitude scale to produce a new scale with even higher reliability.

A correlation matrix for all items and attitude constructs (i.e. the mean of the scores on the items grouped together) was calculated, and whenever the correlation between an item and a scale in which it was not included was high enough to suggest that its inclusion might increase the alpha reliability of the scale, the alpha value was recalculated with the item included. None of the scales derived from factor analysis had their internal consistency (coefficient alpha) increased by the inclusion of any other item. However, small increases in the consistency of some of the 'face validity' constructs were achieved. For example, the alpha value of the 'Attitude to ALIS' construct was increased (from 0.77 to 0.80) by the inclusion of B29 ('I think the Head/Principal should not use ALIS results in staff appraisal'; inverse coded). On reflection, it seemed that this item could well have been included in the original 'Attitude to ALIS' group based on face validity, since it could be interpreted as indicating endorsement of the value of ALIS results. It was therefore decided to include it in that construct. Other small increases were achieved by adding B21 (inverse coded) to 'Self-Efficacy (alpha from 0.78 to 0.82), adding B20 (inverse coded) to 'Achievement Orientation' (alpha from 0.79 to 0.82), and adding B14 (inverse coded) to 'Attitude to ALIS' (alpha from 0.77 to 0.82). However, in none of these cases was it felt that the new item brought the meaning of the construct closer to what had originally been intended, and they were not subsequently included.

Synthesis and overview of attitude scales

The results of all these different methods of constructing attitude scales seem at first sight to be rather hard to integrate. Broadly speaking, two methods (face validity and factor analysis) produced a set of constructs each, while the other methods either

failed to produce satisfactory scales (cluster analysis) or made only small changes to the scales already found (Likert method). The attitude scales formed are summarised in Table 10.

Table 10: Summary of attitude scale constructs

SOURCE	CONSTRUCT	COMPONENT ITEMS	NO. OF ITEMS	ALPHA
Face Validity	Self-Efficacy	B04, B14i, B15i, B17, B20i, B25, B26i	7	0.78
	Achievement Orientation	B01, B02, B05i, B06, B14i, B21i, B26i, B29i	8	0.79
	Locus of Control	B03, B12, B17, B22	4	0.83
	Attitude to ALIS	B06, B10i, B13, B19i, B20i, B26i, B29i	7	0.80
Factor Analysis	Feedback Anxiety	B14, B20, B21, B26	4	0.85
	Responsibility	B03, B12, B17, B22	4	0.83
	Self Confidence	B04, B25	2	0.84
	Feedback Desire	B01, B02, B06, B26i	4	0.74
	ALIS Value	B10i, B19i	2	0.79
	ALIS Fairness	B06, B13, B29i	3	0.69

Note: Items followed by 'i' are inverse coded.

The two sets of constructs do, however, have some features in common. Firstly, and most obviously, examination of their constituent items shows that 'Locus of Control' is identical to 'Responsibility'. The items in these scales are measuring the extent to which people perceive themselves to have control over (and are therefore responsible for) their students' performance.[7]

Secondly, a similar examination shows that the items in 'ALIS Value' and 'ALIS Fairness' are all contained in the 'Attitude to ALIS' scale. The latter may thus be thought of as incorporating two distinct but related components: a person's generally positive attitude towards ALIS, for example, might be expected to indicate that they see it as both of some value and a fair measure of performance. McKennell's (1977) advice for a situation where a construct can be spilt into components is that, although in general the components may be highly correlated which makes it tempting

[7] In order to avoid duplication, 'Responsibility' was therefore dropped from any further analysis.

to combine them into a single variable, there may be a sub-sample for whom they are not correlated and it is therefore of interest to retain the component parts.

Continuing in this vein, it can be seen that 'Self-Confidence' is wholly contained within 'Self-Efficacy', and similarly 'Feedback Desire' is contained in 'Achievement Orientation'. Both of these inclusions seem intuitively reasonable. The only 'factor analysis' factor not to be contained within a 'face validity' factor is 'Feedback Anxiety'. This factor has considerable overlap with both 'Self-Efficacy' and 'Achievement Orientation' (its component items being reverse coded in these). Once again, this overlap has a high level of plausibility. Those who have a high perception of their own efficacy might be expected to be less anxious about receiving performance feedback, as might those who tend to place a high value on achievement and performance feedback. The inter-relationships among the various attitude constructs are shown further in a matrix of their inter-correlations (Table 11).

It therefore seems that the effect of factor analysis is to split broad factors into finer, more uni-dimensional sub-factors. Both kinds of constructs are of value: the former because they capture a broad intuitively based impression of particular relevant attitudes, each derived from a relatively large number of items, and with significant overlaps between them; the latter because they are more strictly uni-dimensional, without overlap of constituent items, but each consisting of fewer items and therefore perhaps more sensitive to particular nuances of wording or context.

Table 11: Inter-correlations among attitude constructs

	SELF-EFFIC'Y	ACHNT ORNTN	LOC OF CTRL	ATT TO ALIS	ANXTY	SELF-CONF.	F'BACK DESIRE	ALIS VALUE	ALIS FAIR
SELF-EFFIC'Y		*l*=0.54 *u*=0.79	*l*=0.14 *u*=0.54	*l*=0.46 *u*=0.74	*l*=-0.90 *u*=-0.77	*l*=0.42 *u*=0.72	*l*=0.26 *u*=0.62	*l*=0.10 *u*=0.51	*l*=0.16 *u*=0.55
ACHNT ORNTN	**0.68** n=72		*l*=-0.12 *u*=0.33	*l*=0.63 *u*=0.83	*l*=-0.88 *u*=-0.71	*l*=-0.05 *u*=0.39	*l*=0.80 *u*=0.92	*l*=0.13 *u*=0.54	*l*=0.55 *u*=0.79
LOC OF CTRL	**0.36** n=72	**0.11** n=72		*l*=0.00 *u*=0.43	*l*=-0.30 *u*=0.15	*l*=0.05 *u*=0.47	*l*=-0.06 *u*=0.38	*l*=-0.11 *u*=0.34	*l*=0.09 *u*=0.50
ATT TO ALIS	**0.62** n=72	**0.75** n=72	**0.23** n=72		*l*=-0.81 *u*=-0.59	*l*=-0.18 *u*=0.27	*l*=0.55 *u*=0.79	*l*=0.67 *u*=0.86	*l*=0.66 *u*=0.85
ANXTY	**-0.85** n=72	**-0.81** n=72	**-0.08** n=72	**-0.72** n=72		*l*=-0.42 *u*=0.02	*l*=-0.71 *u*=-0.39	*l*=-0.54 *u*=-0.13	*l*=-0.58 *u*=-0.20
SELF-CONF.	**0.59** n=72	**0.18** n=72	**0.27** n=72	**0.05** n=72	**-0.21** n=72		*l*=-0.20 *u*=0.26	*l*=-0.27 *u*=0.19	*l*=-0.17 *u*=0.28
F'BACK DESIRE	**0.46** n=72	**0.87** n=72	**0.17** n=72	**0.69** n=72	**-0.57** n=72	**0.03** n=72		*l*=0.09 *u*=0.50	*l*=0.47 *u*=0.75
ALIS VALUE	**0.32** n=70	**0.35** n=70	**0.12** n=70	**0.78** n=70	**-0.35** n=70	**-0.04** n=70	**0.31** n=70		*l*=0.24 *u*=0.61
ALIS FAIR.	**0.37** n=72	**0.69** n=72	**0.31** n=72	**0.77** n=72	**-0.41** n=72	**0.06** n=72	**0.63** n=72	**0.44** n=72	

Figures in bold (below diagonal) are Pearson product moment correlation coefficients (with number of pairs). Figures above diagonal are approximate lower (l) and upper (u) (95%) confidence limits, calculated using Fisher's Z-transform.

Triangulation: agreement between attitude constructs and attitudes inferred from open comments

The ratings of respondents' attitudes towards ALIS from their open comments (see p107) were combined to give an 'average' rating. Each of four raters had been asked independently to classify the open comments made on the questionnaire as 'generally positive' 'generally negative' or 'neutral/mixed'. These ratings were coded as 1, -1 and 0 respectively and for each questionnaire the mean of the four was calculated. This 'mean attitude rating' had an internal consistency (Cronbach's alpha) of 0.88. The mean was also calculated using all five available ratings, one of which was from the same person again. This scale was found to have a slightly higher value of Cronbach's alpha (0.92). Both of these values are high enough to suggest that combining the individual attitude ratings produced a measure with very acceptable reliability, despite the fact that, taken individually, the attitude ratings showed considerable variation.

Correlations between both the 4-rater mean and the 5-rater mean and each of the attitude constructs derived from the Likert items are shown in Table 12. Given that

the mean attitude rating from the open ended comments might have been expected to be measuring much the same as the construct 'Attitude to ALIS', it is somewhat surprising – and perhaps a little disappointing – that the correlation is not higher than the value 0.46 that was found. As a test of concurrent validity, a correlation coefficient of this size is not very impressive. Either the two variables are measuring different things, or they are measuring the same thing not very accurately. Taking the internal consistency (Cronbach's alpha) of each measure as an estimate of its reliability, we can make a correction for attenuation,[8] that is to say an estimate of what the correlation would have been if both variables had been measured with perfect reliability. This correction raises the above coefficient to 0.54. It is interesting that the mean attitude rating's highest correlation (0.54) is with 'ALIS Fairness' (this value becomes a slightly more respectable 0.68 when corrected for attenuation). Whether this is because the aspects of people's attitudes that were identified by the raters as 'positive' of 'negative' were particularly concerned with perceptions of fairness, or whether it simply indicates the complexity of the attitudes involved is hard to say. It is certainly arguable that the difference between the two coefficients is not large enough to warrant attention.

[8] The 'corrected' estimate is given by $r_{xy}/\sqrt{(r_{xx}.r_{yy})}$, where r_{xy} is the measured correlation coefficient and r_{xx} and r_{yy} are the reliabilities of the two variables (McKennell, 1977). It is potentially somewhat misleading, however, merely to cite the corrected coefficient, since the effect of unreliability in the two measures will not only reduce the maximum correlation between them, but also increase substantially the amount of potential error in the estimate. The correction for attenuation provides a maximum likelihood estimate of the true correlation, but without giving any indication of how much the confidence interval around it has been increased.

Table 12: Correlations between attitude towards ALIS from open comments and attitude constructs

CONSTRUCT	CORRELATION WITH	
	Mean attitude rating (4 raters)	Mean attitude rating (5 raters)
Self-Efficacy	0.09	0.12
Achievement Orientation	0.21	0.23
Locus of Control	0.12	0.13
Attitude to ALIS	0.44	0.46
Feedback Anxiety	-0.08	-0.10
Responsibility	0.12	0.13
Self Confidence	0.02	0.03
Feedback Desire	0.23	0.25
ALIS Value	0.38	0.40
ALIS Fairness	0.53	0.54

n = 72

The fact that the mean attitude rating's largest correlations were with the three constructs concerned with attitudes towards ALIS is to some extent an endorsement of their previous interpretations. Certainly, the relative sizes of the coefficients are consistent with what might have been predicted, and these results may therefore be seen as supporting the 'construct validity' of these constructs (Kerlinger, 1986). The absolute sizes are, however, a little disappointing. Perhaps the safest conclusion from the triangulation attempt is that the constructs may well be broadly measuring what they were intended to measure, but the underlying attitudes could be more complex than was supposed, and there is also a good deal of noise around the signal.

Implications for the design of the final questionnaire

The final questionnaire was intended to provide a post-test measure of the same attitudes as measured in the initial questionnaire. It was therefore necessary for it to contain all the items that had contributed to any of the attitude scales constructed from the above analysis, and equally, there was no need for it to include any of the others. It was thus possible to remove items B08, B09, B11, B16, B18, B23, B27, and B28.

6.3 IDENTIFICATION OF TEACHING GROUPS

Once the information about teaching groups had been received, the teachers in the sample were allocated to either 'treatment' or 'control' groups (see Chapter 5). However, this information was supplied by only some of the departments, so the sample used in the experiment was significantly smaller than (and potentially quite different from) the original sample. It was therefore necessary to repeat some of the exploratory data analysis described in Section 6.2 above.

Characteristics of the experimental sample

The sample used for the experimental intervention consisted of 44 teachers in six institutions, although all but nine of them were in three of the institutions (see Table 13). It would be fair to say, therefore, that the bulk of the experiment was conducted in these three main institutions, a fact that inevitably reinforces even further concerns about representativeness and generalisability of the results.

Table 13: Institutions represented in the experimental sample

INSTITUTION	NUMBER OF TEACHERS		
	Control group	Treatment group	Total
Inst1	3	6	9
Inst3	7	5	12
Inst4	7	7	14
Inst6	2	3	5
Inst7	2	1	3
Inst9	1	0	1
Total	22	22	44

The number of teachers in each 'subject' and 'position in the institution' category was also calculated, since it was thought possible that either of these variables might interact significantly with the effects of the feedback given, and it was therefore important to know whether each of the different subjects had broadly the

same pattern of positions of responsibility. If this were not the case, it would be easy to confuse the effect of one of these variables for that of the other.

It was felt to be necessary to group some of the categories together, partly because of the small number of people in the sample,[9] and partly for the sake of clarity. Fortunately, analysis of the later questionnaires (see below) suggested that there might be important differences between the way information was treated by those with a 'numerical' background (i.e. teachers of mathematics and physics) and those with a 'non-numerical' background (French and English), so the subjects were grouped this way.

Table 14: Subject type and position of experimental sample

SUBJECT TYPE	POSITION			
	Subject Teacher	Department Responsibility	Management Responsibility	Total
Numerical (Maths, Physics)	12	9	3	24
Non-numerical (English, French)	9	5	4	18
Total	21	14	7	42

Note: The 'Position' of two respondents was classified as 'other'. These have been excluded from this analysis.

It can be seen from Table 14 that the spread of positions held is comparable for both subject types. A chi-squared test confirmed that there was no significant interaction ($\chi^2 = 0.875$, $p = 0.65$).

6.4 IMPLEMENTATION-CHECK QUESTIONNAIRE

This questionnaire was sent to those who had recently received the feedback to find out how much time they had spent (or would spend) on it, and how useful and easy to understand they had found it. The responses were all pre-coded and the frequencies of each response are given in Appendix 6I (p292). In addition,

respondents were invited to comment on the feedback and the comments made were transcribed and are presented in Appendix 6J (p293).

Summary of responses

Responses were received from 15 of the 22 teachers who received the feedback (i.e. 68%). Thus, about a third did not reply and inferences must be treated with appropriate caution. Most people reported spending between 5 and 20 minutes on each part of the feedback, and most intended to spend less than 5 minutes further. One comment made (see response 5, Appendix 6J, p293) suggested that more time might have been spent if the data had related to more than 8 students. Another (from a lecturer in an FE college, response 13) explained the pressures and tensions within his institution, with the implication that, had it not been for these, he might have spent more time on it and been able to answer the questions he left blank. A large majority (11) said they had discussed the feedback with colleagues. The only comment made relating to this issue was by a head of department who said that he would have like to have seen the information for each of the members of his department (response 8).

The question asking 'How easy to understand have you found it?' produced perhaps the most interesting responses. A full range of views was achieved, with three people finding it 'very easy' but two saying it was 'impossible'. This seems to suggest that what to one person may have been a clear and transparent numerical summary of their students' performance, to another may have been a meaningless jumble of figures. This difference in ease of understanding seemed to be closely related to the subject taught, and the issue of subject differences is discussed below. On balance, the majority found the feedback accessible, with ten people saying it was either 'easy' or 'very easy' to understand, against three who found it 'hard' or 'impossible'. Comments made by two of those in the latter category (both teachers of English) drew attention to the difficulties of interpreting the ideas of 'significance' and 'deviation' (response 11) and to the need for a more 'user friendly' form of presentation (response 7).

---

[9] This is an issue in, for example, the χ^2 test, where it is generally held that a χ^2 value based on a contingency table containing expected frequencies of less than 5 is hard to interpret (see footnote 11, p136).

The parts of the feedback which were reported to be the most useful were the individual student results ('Student results') and the overall summary of all the results for that teacher over the four years ('Summary by teacher'), each of which were categorised by seven respondents as either 'useful' or 'extremely useful'. Even this, however, represents only half of the respondents[10] having rated it as at least 'useful'. The least useful part was the target grades for 1997 ('Targets 97'), with only two respondents attributing it the same value. One comment made (response 8) explained the lack of use of the latter as being a result of its timing (too late), its lack of accuracy (especially compared with teacher judgement) and doubts about the appropriateness of using ALIS data in this way.

Taking the feedback as a whole, the respondents can be divided into those who rated most of it (i.e. three or more of the five parts) to be 'useful' or better, and those who did not. With this classification, five found it broadly useful and nine did not. Among the nine, however, were the four respondents who found it 'moderately difficult' or harder to understand, and the one who had had only eight students. Comments which related to the general perception of usefulness of the feedback included the statement that they had already done a similar analysis for themselves (response 2), a comment that it would have been 'tremendously useful' in the Autumn (response 8) and the remark that it was useful, but not essential (response 9).

Finally, two comments were made which seemed to reflect perceived shortcomings in the validity of the ALIS model. One suggested that some of the 'predicted' grades were unrealistically optimistic, commenting 'I'm not a believer' (response 1). Another drew attention to the failure to take account of student absence as an explanation for performance (response 11).

Subject differences

Although the number of respondents was so small, there were nevertheless some interesting differences between the replies of teachers of different subjects. In particular, it was noticed that all of those who had described the feedback as 'very easy' or 'easy' to understand taught either mathematics or physics, and all who had

[10] One of the 15 who returned the questionnaire (response 13) commented that he had not studied the information sufficiently and left questions 5 and 6 blank. These questions were therefore treated as

said it was 'moderately hard', 'hard' or 'impossible' to understand had been teachers of English or French. Although it might have been expected that teachers of the more numerical subjects would generally find information that consisted of statistics and graphs easier to make sense of, such a clear separation of the two groups was not anticipated. As these ratings of 'ease of understanding' may be considered to form an ordinal scale, a Mann-Whitney U-test was used to establish the statistical significance of such an extreme split. A 'p' value of 0.002 was obtained, which suggests that it is extremely unlikely that such a difference would have arisen by randomly sampling from a population of teachers for whom their subject was unrelated to their ease of understanding. It therefore seemed appropriate to analyse the responses separately for teachers of 'numeric' subjects (Mathematics and Physics) and 'non-numeric' subjects (English and French).

At first sight, it appears highly significant that of the 15 replies received, 11 were from teachers of 'numeric' subjects and only 4 from 'non-numeric'. However, the two groups were not equally represented in the treatment group: the 22 teachers who received the feedback comprised 14 of 'numeric' subjects and 8 'non-numeric'. A chi-squared test for the independence of subject type and questionnaire response with these frequencies gives $p = 0.17$.[11] Hence it would not be at all unlikely for such a difference in response rates to have been a result of pure chance.

Having established that the two subject types were different with respect to how easy they found the feedback to understand, they were then compared to see whether there were differences in any of the other questions. On the Mann-Whitney U-test none of the other questions showed statistically significant differences between the subjects at the 5% level, though the ratings of the usefulness of the graphs came

---

having 14 responses.

[11] The contingency table of expected values for these data does have two values below 5 and therefore violates a commonly adopted condition for the use of a chi-squared test. However, other authorities argue that this is unnecessarily restrictive. Quadling (1987, pp86-7, 331-2) advises caution in interpretation of χ^2 values where either low expected frequencies may make the test statistic particularly sensitive to small changes in observed frequencies, or where one or more of the observed frequencies differs very markedly from the expected frequency (say by more than about 50%). In the case under consideration here, the latter objection does not apply, although the former may do. Camilli and Hopkins (1978, 1979) argue that the test is reliable provided the *average* expected frequency is at least 2, which would make its use in this case perfectly acceptable.

There is also controversy about whether Yates' correction for continuity should be made. Camilli and Hopkins (1978) argue that it is unnecessary and distorts already conservative alpha values to be more so. Its use in this case would have given $p = 0.36$, and would therefore not have altered the conclusion.

extremely close with p = 0.053.[12] All four of the non-numeric respondents had rated the graphs 'of some use', as had four of the numeric group, while a further five of the latter had rated it 'useful' and one 'extremely useful'

Relationships among variables

A summary statistic for each of the three main pieces of information collected by the questionnaire (time spent, ease of understanding and usefulness) was calculated as follows:

The time spent (already and expected) was coded for each of the two parts of the feedback as 'less than 5 mins' = 1, '5-20 mins' = 2, '20 mins-1 hr' = 3, 'more than 1 hr' = 4, and the total of these four codings denoted by 'TIME'. 'EASE' (of understanding) was coded as 'very easy' = 5, 'easy' = 4, 'moderately hard' = 3, 'hard' = 2, 'impossible' = 1. The usefulness rating for each of the five parts of the feedback was coded as 'extremely useful' = 4, 'useful' = 3, 'of some use' = 2, 'no use at all' = 1, and the total of the five codings denoted by 'USE'.

Correlations were calculated for each pair, none of which were found to be significant, given only 14 responses. In fact, no correlation had absolute value above 0.24. Examination of the appropriate scatter graphs confirmed the absence of any association. Thus, for this small sample at least, there was no apparent relationship between the amount of time a person spent (or intended to spend) on the feedback, how easy they found it to understand and their perception of its usefulness.

There were, however, some interesting relationships between respondents' perceptions of the feedback (as measured by 'EASE', 'TIME' and 'USE') and the content of the information it contained. The content of the feedback was summarised by three variables: 'STD_RES', the mean standardised residual gain of the students taught by that teacher, 'REL_VA', the relative value added (i.e. the difference between the mean of the students' standardised residuals in the teacher's subject and the same students' overall performance in all their subjects), and 'ATTITUDE', the

[12] In fact, giving this probability to even one decimal place may be overestimating its accuracy (see the section in Chapter 4 on significance tests), so quibbling over the third decimal place is arguably quite absurd. However, the logic of significance testing requires that an arbitrary cut-off be applied, without judgement.

mean of the students' ratings on the ALIS 'attitude to subject' scale. The relevant correlations (together with lower and upper confidence limits) are shown in Table 15.

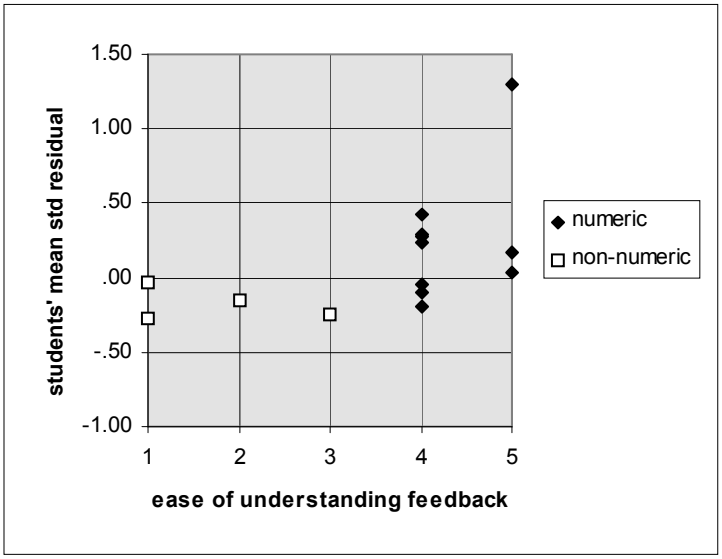
Table 15: Correlations between content of feedback and how perceived

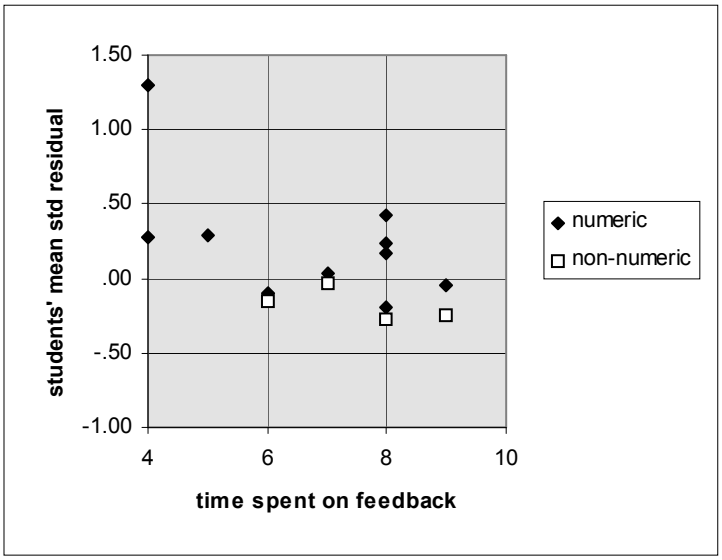
STUDENTS' MEAN:	PERCEPTION OF FEEDBACK								
	Ease of understanding			Time spent			Usefulness rating		
	r	n	95% C.I.	r	n	95% C.I.	r	n	95% C.I.
Standardised Residual	**0.53**	14	[0.06, 0.80]	**-0.58**	15	[-0.82, -0.16]	0.48	14	[0.00, 0.78]
Relative Value Added	**0.50**	14	[0.03, 0.79]	**-0.49**	15	[-0.78, -0.03]	0.46	14	[-0.02, 0.77]
Attitude to Subject	0.15	14	[-0.36, 0.59]	-0.25	15	[-0.64, 0.25]	0.36	14	[-0.15, 0.71]

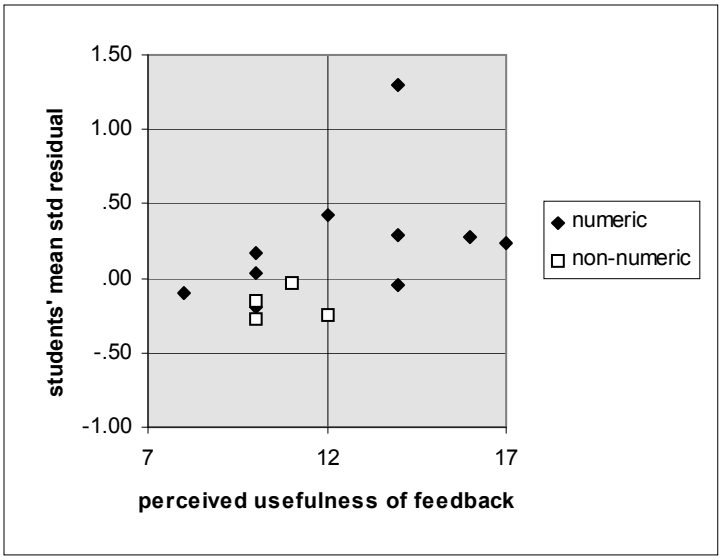
r = product moment correlation (bold indicates statistically significantly different from 0); n = number of pairs; 95% confidence interval derived from Fisher Z-transform.

From Table 15 it appears that the people whose feedback was most positive (in terms of student performance, measured by both standardised residuals and relative value added) tended to report finding the feedback easier to understand and also spent (or intended to spend) less time on it. The correlation between students' performance and the overall perception of the usefulness of the feedback was also positive and of the same magnitude (around 0.5), but just too low to be considered statistically significant at the 5% level. The correlations between students' attitudes and these three variables were in the same directions, but appreciably lower. However, correlations derived from such a small sample must be interpreted cautiously, even if they are 'statistically significant'. Scatter graphs for the three correlations with standardised residuals are shown in Figure 5, with teachers separated by subject type.

Figure 5: Scatter graphs of student performance with 'ease', 'time' and 'use'







The overall impression gained from the graphs in Figure 5 is of a slight association, but one which could easily be quite dependent on a small number of crucial cases.

6.5 FINAL QUESTIONNAIRE

Likert items

For each of the Likert items, the correlation between a person's scores on the two questionnaires was calculated. The value of this when restricted to the control group is generally cited as the test-retest correlation for that item. The correlations for the feedback group and for all the respondents together were also calculated for comparison (Table 16). The number of respondents in each of the treatment groups whose scores were the same on both questionnaires, the number whose scores had gone up (i.e. who agreed more strongly with the statement on the final questionnaire than they had on the initial), and the number whose scores had gone down (i.e. who disagreed more strongly with the statement on the final questionnaire than they had on the initial) were also calculated for each item (Table 16).

Table 16: Test-retest correlations and changes in responses for Likert items on final and initial questionnaire

ITEM	TEST-RETEST CORRELATIONS			NUMBER OF CHANGES					
	Sample restricted to:			Feedback			Control		
	All	Feedback	Control	Same	Up	Down	Same	Up	Down
B01 / A	0.60	0.60	0.64	11	5	3	12	8	1
B02 / B	0.57	0.46	0.66	7	6	6	13	5	3
B03 / C	0.55	0.51	0.64	9	3	7	12	4	5
B04 / D	0.66	0.72	0.63	14	3	1	14	2	4
B05 / E	0.59	0.50	0.66	8	6	5	7	9	4
B06 / F	0.34	0.37	0.53	8	3	7	10	8	3
B10 / G	0.56	0.32	0.72	6	5	7	11	1	8
B12 / H	0.51	0.66	0.38	8	7	4	11	5	5
B13 / I	0.53	0.49	0.65	9	2	7	13	7	1
B14 / J	0.48	0.50	0.45	8	3	7	7	4	10
B15 / K	0.69	0.69	0.72	10	5	3	10	5	5
B17 / L	0.63	0.80	0.64	12	0	6	10	8	3
B19 / M	0.38	0.06	0.56	7	4	6	7	4	9
B20 / N	0.60	0.64	0.55	9	4	5	11	5	5
B21 / O	0.59	0.57	0.66	7	4	6	7	10	4
B22 / P	0.64	0.73	0.62	12	3	4	7	11	3
B25 / Q	0.66	0.66	0.75	16	3	0	16	1	4
B26 / R	0.56	0.66	0.56	8	5	4	11	3	7
B29 / S	0.81	0.88	0.70	14	1	2	12	5	3

Note: 'Up' = number of respondents who agreed more strongly on final than initial questionnaire

It can be seen from Table 16 that the (control group) test-retest correlations for the Likert items are generally satisfactory, with an average value of 0.62. The lowest two correlations are just 0.38 (for item B12/H) and 0.45 (B14/J), which suggest that the responses to those items were not very stable, but the removal of a single outlier in each case can improve the control group correlations to 0.57 and 0.67 respectively.

From the figures for the changes in response (Table 16) it can be calculated that slightly over half (52%) of the responses on the final questionnaire were exactly the same as they had been on the initial version. Overall, this was consistent across treatment groups, the percentage the same being 53% and 51% for those in the feedback and control groups, respectively. Of the responses that had changed, overall, equal numbers had changed in each direction, and there was no significant difference between the feedback and control groups.

Attitude constructs

Construct reliabilities

The reliability of a measuring instrument such as a questionnaire is generally defined in two ways: either by some measure of internal consistency (e.g. split half or Cronbach's alpha), or as a test-retest correlation.

The internal consistency (Cronbach's alpha) of each of the attitude constructs had already been calculated for the sample replying to the initial questionnaire (see Table 10, and reproduced below). These calculations were repeated for the responses to the final questionnaire in order to check that the items related to each other in the same way as they had done on the initial questionnaire. Test-retest correlations were also calculated for each of the constructs, based on the two sets of responses of all those in the control group.

Table 17: Reliabilities of questionnaire attitude constructs

CONSTRUCT	RELIABILITY		
	Alpha from initial questionnaire (all returns, n ≈ 72)	Alpha from final questionnaire (n ≈ 40)	Test-retest correlation (ctrl group, n ≈ 21)
Self-Efficacy	0.78 [0.71, 0.85]	0.72 [0.55, 0.81]	0.50 [0.12, 0.75]
Achievement Orientation	0.79 [0.68, 0.83]	0.81 [0.69, 0.88]	0.79 [0.57, 0.90]
Locus of Control	0.83 [0.73, 0.90]	0.81 [0.67, 0.89]	0.79 [0.57, 0.90]
Attitude to ALIS	0.80 [0.74, 0.86]	0.82 [0.69, 0.91]	0.61 [0.27, 0.81]
Feedback Anxiety	0.85 [0.78, 0.90]	0.80 [0.68, 0.90]	0.53 [0.16, 0.77]
Self Confidence	0.84 [0.74, 0.91]	0.76 [0.41, 0.89]	0.77 [0.53, 0.90]
Feedback Desire	0.74 [0.60, 0.84]	0.72 [0.54, 0.82]	0.73 [0.46, 0.88]
ALIS Value	0.79 [0.65, 0.91]	0.54 [0.08, 0.78]	0.69 [0.39, 0.86]
ALIS Fairness	0.69 [0.52, 0.80]	0.82 [0.73, 0.89]	0.69 [0.40, 0.86]

Note: 95% confidence intervals [shown in brackets] are derived from bootstrapping (see Annex) for alpha, and Fisher Z-transforms for correlations. Values of 'n' indicate maximum numbers; confidence intervals are based on actual numbers of responses.

Overall, these indicators of reliability are fairly good. The internal consistencies of the constructs appear to have held up reasonably well on the repeated questionnaire. Although six of the nine alpha coefficients have dropped, this 'shrinkage' would have been expected, given that the constructs were selected in large part for their high values of alpha with the original sample. Inevitably, part of their apparent 'consistency' on the initial questionnaire will have been particular to that sample. Moreover, the drop in alpha is in all but one case ('ALIS Value') too small to be considered significant. The absolute values of alpha themselves also compare well with what might have been expected. A summary by Stipek and Weisz (1981) of the 11 most commonly used instruments for measuring 'Locus of Control' found reported values of Cronbach's alpha between 0 and 0.87, but the average of the values they report is just 0.56.

Test-retest correlations were also mostly satisfactory. For six of the constructs they were above 0.67, the average value reported by Stipek and Weisz (1981) for measures of 'Locus of Control'. This sort of value is also comparable with published 12-week test-retest correlations for other attitude tests (e.g. Viswanatham, 1994; Van Ryckeghem and Brutten, 1992). Considering that the two questionnaires were answered more than six months apart, this seems to indicate that these attitudes (as measured by this questionnaire) were reasonably stable. Two of the correlations ('Self-Efficacy' and 'Feedback Anxiety') were closer to 0.5, which is a little lower than might have been hoped,[13] and may indicate a need for caution in any interpretation of apparent changes.

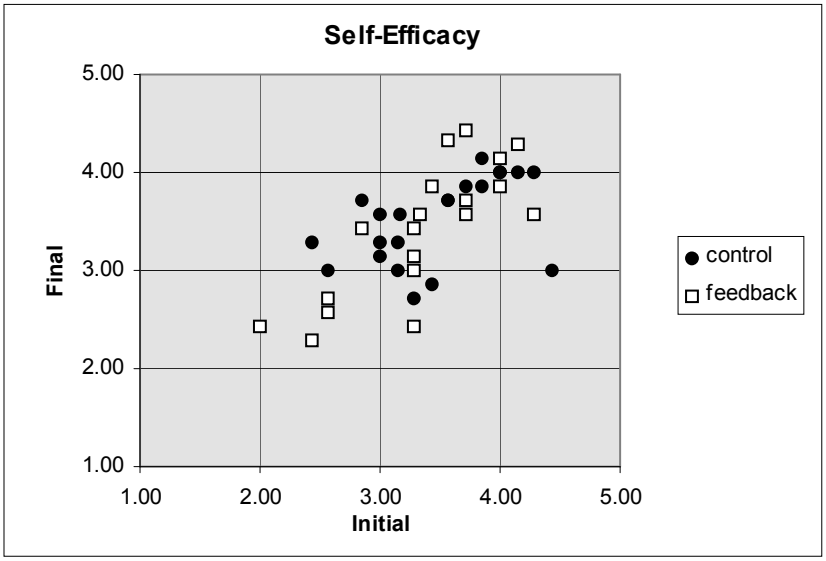
Changes on constructs: Scatter graphs

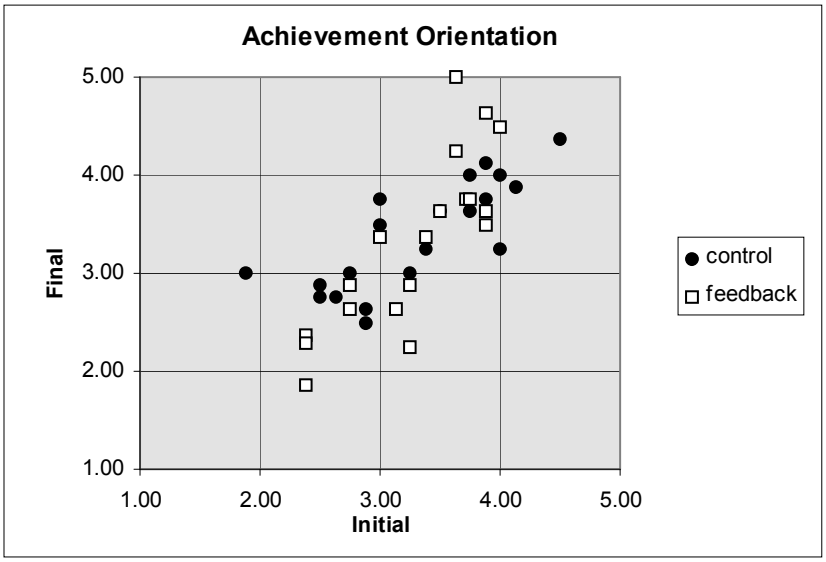
No statistical test provides as much information as an examination of the original data, and a scatter graph of the distribution of responses on the initial and final questionnaires shows clearly not only the nature of test-retest relationship, but also the extent of the differences between the respondents who received the feedback

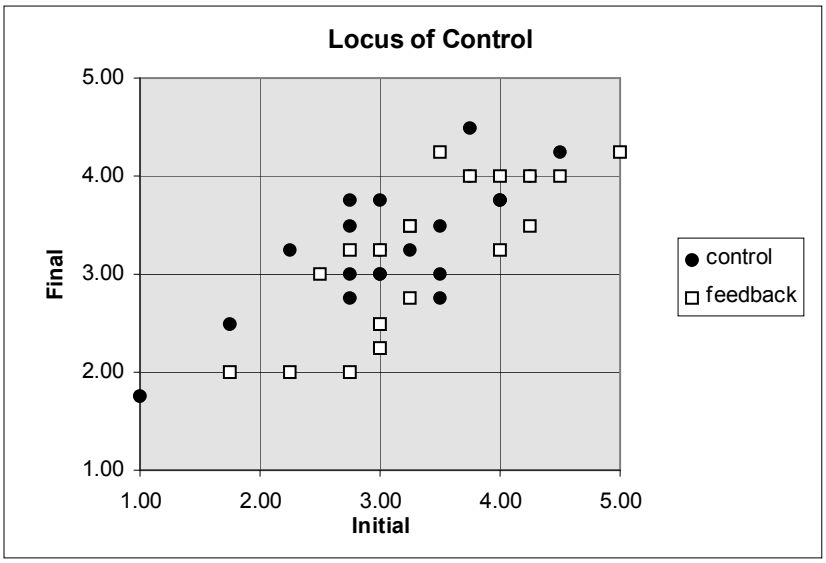
[13] After plotting scatter graphs of the scores on the two questionnaires (see Figure 6), it was found that both these low correlations could be improved substantially by the removal of a single outlier. For 'Self Efficacy' the point (4.43, 3.0) was removed, raising the correlation to 0.69, and for 'Feedback Anxiety' the point (1.25, 3.5) was removed to give a correlation of 0.73. Interestingly, both points represent the same person. These examples illustrate the sensitivity to outliers of correlation coefficients calculated from small (here $n \approx 20$) samples.

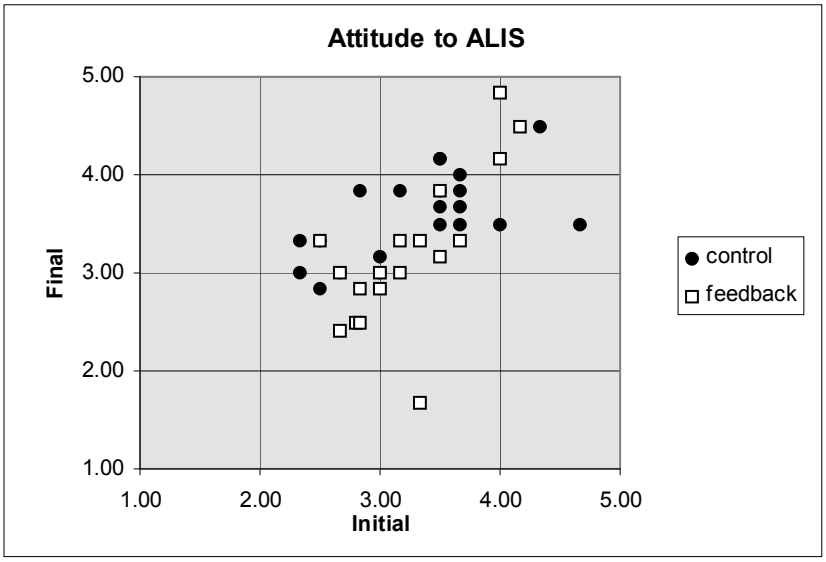
and those in the control group. Scatter graphs for each construct are shown in Figure 6.

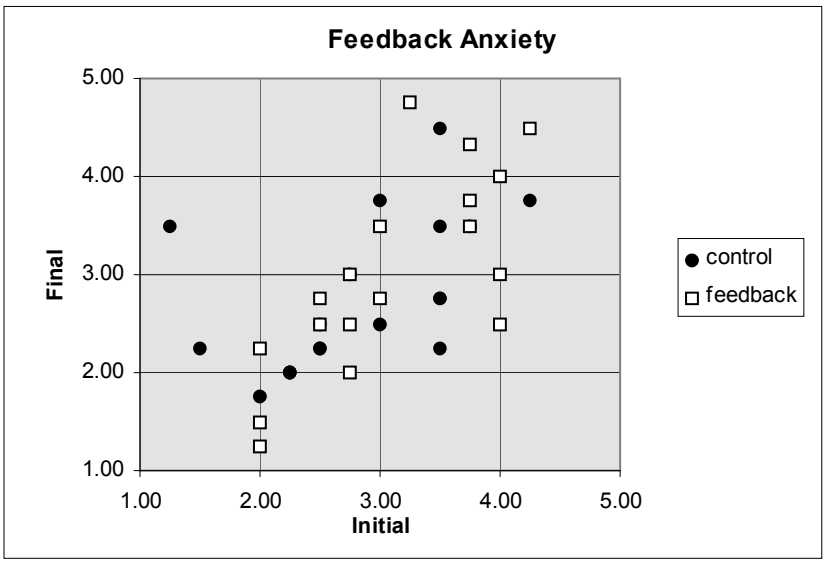
Figure 6: Scatter graphs of initial and final scores for each attitude construct.



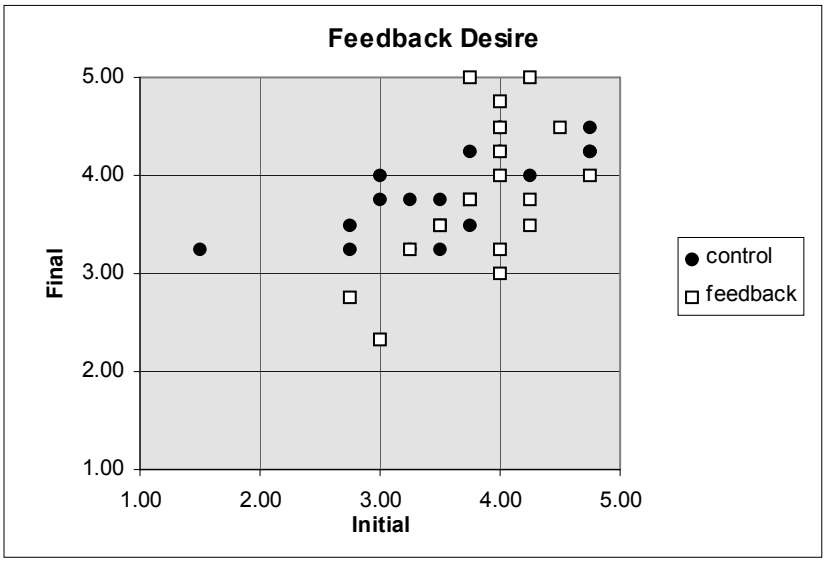


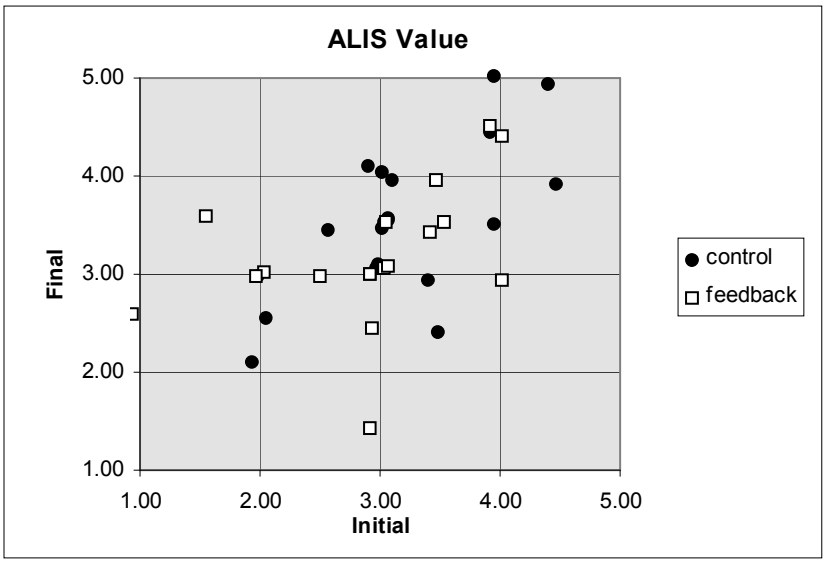

Locus of Control
Final
Initial
5.00
4.00
3.00
2.00
1.00
1.00
2.00
3.00
4.00
5.00
control
feedback

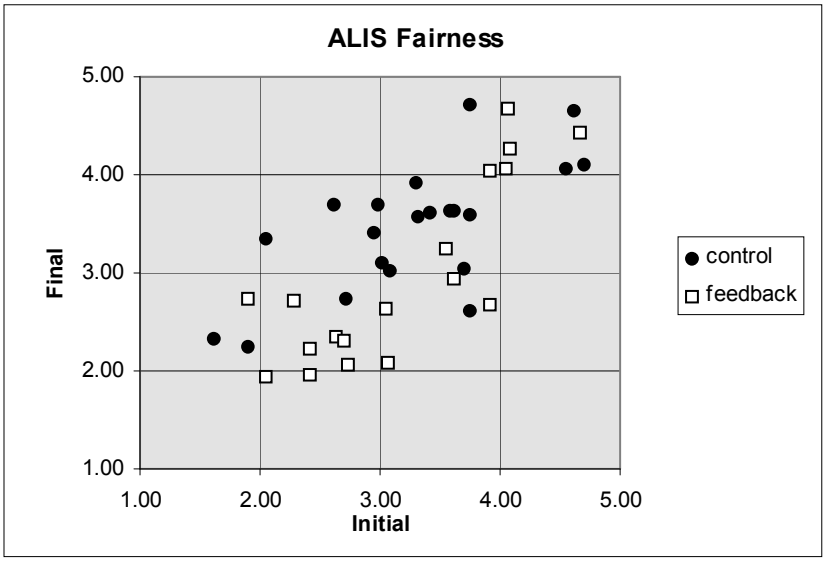

Attitude to ALIS
Final
Initial
5.00
4.00
3.00
2.00
1.00
1.00
2.00
3.00
4.00
5.00
control
feedback


Feedback Anxiety
Final
Initial
5.00
4.00
3.00
2.00
1.00
1.00
2.00
3.00
4.00
5.00
control
feedback


Self Confidence
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
control
feedback


Feedback Desire
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
control
feedback





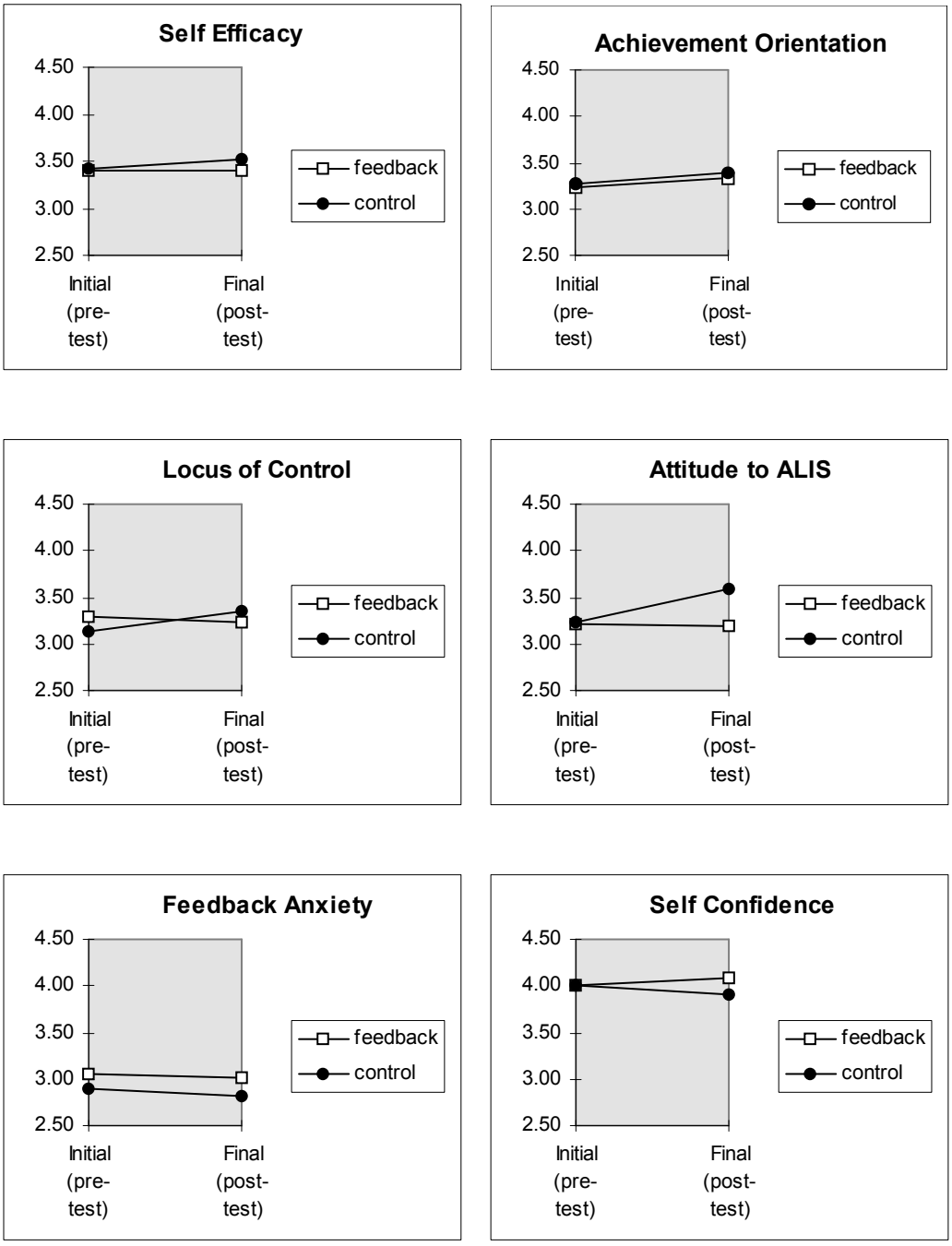
Note: For constructs with fewer than four component items (i.e. Self Confidence, ALIS Value, ALIS Fairness), points have been 'jiggled' by incorporating a small random part to prevent them being coincident.

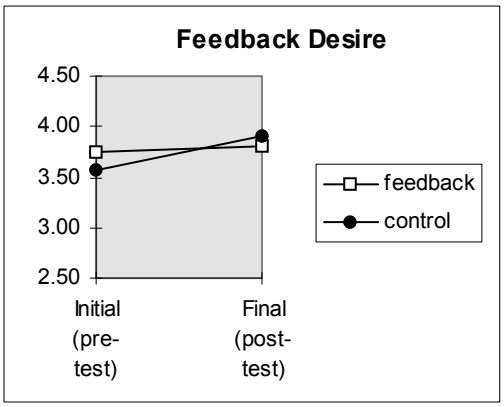
The most striking feature of the graphs in Figure 6 is that the control and feedback groups do not appear to be significantly different on any of them, with the possible exception of the last, 'ALIS Fairness'. Although there are some differences, the graphs illustrate well the variability found with a small sample such as this.

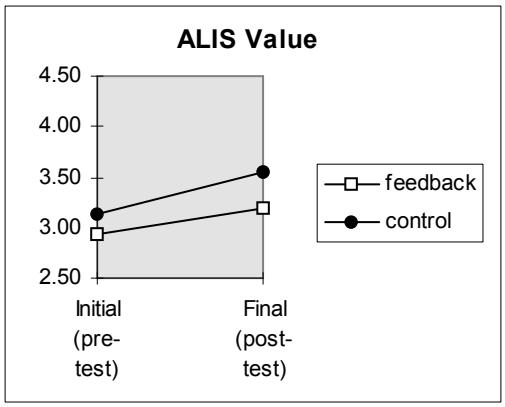
Changes on constructs: Attitude scale means

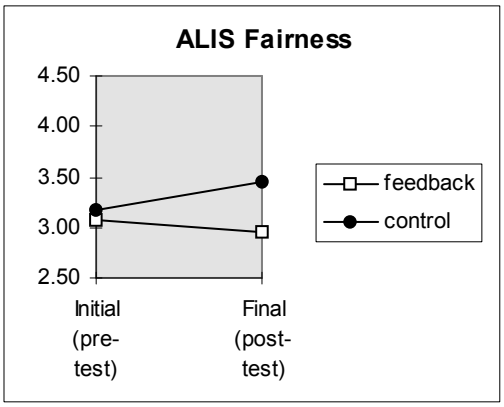
For each attitude construct, the mean score was calculated for both feedback and control groups on both questionnaires. Figure 7 shows the change for each graphically. The exact values for each mean can be found in Table 20 (page 154).

Figure 7: Changes in attitude construct means









Once again, the differences between the feedback and control groups do not seem to be very large, apart from possibly on 'Attitude to ALIS' and 'ALIS Fairness'. On these two attitudes, however, it is the control group that apparently became more positive towards ALIS, while the group who received the feedback became less positive.

Changes on constructs: Absolute Change, Residual Gain or Raw Post-test scores

The main purpose of the final questionnaire was as a post-test measure of the attitude constructs established by the (pre-test) initial questionnaire. In particular, it was hoped to produce an indicator for the change on each construct, in order to quantify the difference between those who received the feedback and those who did not. However, the calculation of such an apparently simple indicator of the effect of the intervention on attitudes is far from straightforward, and the concept of 'change' is somewhat problematic.

As each attitude was measured by the same instrument on both occasions, it is possible to define the change simply as the difference between the two scores. This

method gives equal weight to both pre- and post-test and defines the change in absolute terms. It has the advantages of being clear and readily interpretable. However, it is also generally recognised to have low reliability (since calculating the difference of two unreliable measures compounds the problem of reliability), and to be negatively correlated with the initial measure (again, as a result of unreliability and regression to the mean). Despite this, Rogosa *et al.* (1982) argue that calculated (sample) difference scores are in fact an unbiased estimate of 'true' (population) differences, and are therefore a valid measure of change. The sample correlation between change score and initial status, on the other hand, is a significantly biased estimate of the true value and should not be taken too seriously. They also show that the reliability of the difference is often not significantly less than the reliability of either component (provided there is sufficient variation in change scores), and that this deficiency, too, is 'more illusory than real' (p735).

An alternative approach is to use a regression model to calculate the 'residual' gain, thus defining the 'change' as the difference between the outcome (post-test) measure and what would have been predicted from the pre-test. This method produces a measure of change which is uncorrelated with measured initial status, and makes use of the correlation between pre-and post-test to maximise reliability of the change measure. However, residuals calculated from sample values using OLS regression are biased (depending on true initial status), have low precision (large error variance), are still correlated with (true) initial status and are not robust to outliers. OLS regression is overly sensitive to outliers, whose residuals may therefore make them seem less atypical than is fair, while residuals of typical values are increased. As a result, various modifications to simple OLS residual gains have been proposed for use in the analysis of change scores, all of which, however, are extremely complicated (Rogosa *et al.*, 1982, p739). Despite all their limitations, OLS residual gain scores are still said to be the 'most frequently used in empirical research' (*ibid*., p738). For the analysis conducted here, it was felt that the additional statistical correctness achieved by 'patching up' the simple residuals did not justify the effort involved. Statistical analysis should be an aid to inference, and interpretation is unlikely to be sound unless any manipulations performed on the data are simple and transparent. OLS residuals may not be perfect, but they are good enough, and they are easily understood.

Yet another position on this issue is taken by Cronbach and Furby (1970) who argue that there is no justification for using change scores of either kind. They favour the use of just the outcome (post-test) scores, since if there has been random allocation, this controls for any pre-test differences, and if there has not, then no adequate control can be made. However, it seems rather an extreme position to ignore totally the pre-test data, whatever its limitations, especially given the small size of the sample used here, and the resulting low power of any test for significant differences. It is interesting to note, given the arguments for the use of either 'absolute change' scores or 'raw post-test' scores, that 'residual gain' scores will always lie between the two, and may therefore be thought of as something of a compromise solution. Where the correlation between pre-test and post-test is high, residual gain scores will approximate closely to absolute change scores; where the correlation is low, they will be close to raw post-test scores.[14]

Finally, Rogosa *et al.*'s (1982) conclusion may be noted that sound inferences about change are difficult to make with data from only two time points, and that, ideally, multiwave data are required. This may be the final word on this (otherwise rather equivocal) subject: if you want to know whether a person's attitude has really changed, then you need to measure it a number of times over a period, not just twice.

For the purpose of this analysis, therefore, it was decided to use both absolute change and residual gain change scores, and to compare the outcomes from using each.

For each attitude construct, the (absolute) change between the initial and final questionnaires was calculated for each individual, and the average change of those in the feedback and control groups was compared. A t-test was used to estimate the probability that such a difference would have arisen by chance (see Table 18). These probabilities depend on the assumption that sample values were drawn independently from a population with Normal distribution. However, since all teachers in the same department were allocated to the same treatment group (and teachers in the same department may be more likely than others to share the same attitudes, or to have changed their attitudes in the same way), their attitudes are probably not independent.

---

[14] These statements are strictly true only if absolute change scores and post-test scores are measured on a scale with mean zero. Otherwise, it may be more correct to say that residual gain scores are 'highly correlated' with them.

Hence the probability derived from this test is likely to be an underestimate of the true probability, and low values (i.e. those that suggest a statistically significant difference) should be treated with some caution.

Table 18: Absolute changes in attitudes for feedback and control groups

CONSTRUCT	ABSOLUTE CHANGE IN CONSTRUCT FOR:						DIFFERENCE (fbk – ctrl)					
	Feedback group			Control group				pooled				
	mean	n	var	mean	n	var	mean	var	ν	t	p	
Self-Efficacy	0.07	19	0.18	0.06	21	0.26	0.01	0.22	38	0.05	0.96	
Achievement Orientation	0.04	19	0.28	0.08	21	0.17	-0.05	0.22	38	0.31	0.76	
Locus of Control	-0.14	19	0.26	0.20	21	0.26	-0.35	0.26	38	2.10	0.04	*
Attitude to ALIS	0.12	19	0.71	0.27	21	0.26	-0.15	0.47	38	0.67	0.51	
Feedback Anxiety	0.08	19	1.02	-0.02	21	0.55	0.11	0.77	38	0.38	0.71	
Self Confidence	0.16	19	0.20	-0.12	21	0.17	0.28	0.18	38	1.99	0.05	
Feedback Desire	-0.05	19	0.35	0.29	21	0.30	-0.33	0.32	38	1.81	0.08	
ALIS Value	0.28	18	0.68	0.35	20	0.37	-0.07	0.52	36	0.30	0.77	
ALIS Fairness	-0.19	18	0.28	0.18	21	0.35	-0.38	0.32	37	2.04	0.05	*

* - *Difference statistically significantly different from 0, at 5% level.*

Residual gains were also calculated for each construct, using OLS regression for all values (i.e. control and feedback groups together). Once again, the average gain for feedback and control groups was calculated, and the t-test used to assess the size of the difference (Table 19). Again, the p values may be thought of as underestimates.

Table 19: Residual gains in attitudes for feedback and control groups

CONSTRUCT	RESIDUAL GAIN IN CONSTRUCT FOR:						DIFFERENCE (fbk – ctrl)					
	Feedback group			Control group				pooled				
	mean	n	var	mean	n	var	mean	var	ν	t	p	
Self-Efficacy	-0.02	19	0.19	0.01	21	0.16	-0.03	0.18	38	0.23	0.82	
Achievement Orientation	-0.03	19	0.31	0.02	21	0.14	-0.05	0.22	38	0.33	0.74	
Locus of Control	-0.14	19	0.23	0.13	21	0.16	-0.27	0.19	38	1.90	0.06	
Attitude to ALIS	-0.18	18	0.34	0.15	21	0.15	-0.33	0.24	37	2.04	0.05	*
Feedback Anxiety	0.00	18	0.49	0.00	21	0.43	0.01	0.46	37	0.04	0.97	
Self Confidence	0.13	19	0.13	-0.11	21	0.15	0.24	0.14	38	1.98	0.05	
Feedback Desire	-0.12	19	0.38	0.11	21	0.11	-0.22	0.24	38	1.40	0.17	
ALIS Value	-0.10	18	0.43	0.09	20	0.34	-0.19	0.38	36	0.90	0.37	
ALIS Fairness	-0.22	18	0.29	0.19	21	0.24	-0.41	0.26	37	2.41	0.02	*

* - *Difference statistically significantly different from 0, at 5% level.*

Effect size of changes

The concept of 'effect size' was introduced by Cohen (1969) as a way of quantifying the difference between two groups, rather than simply reporting it to be 'significantly' different from zero. Cohen defines the effect size index, *d*, as the difference in means, divided by the standard deviation (p18). In this context where *d* is an alternative to the t-test, the concept of 'standard deviation' is unproblematic, since the latter is assumed to be equal for both groups. However, with real data, the standard deviations of the two groups are unlikely to be equal, and the assumption that they are sampled from populations with the same standard deviation may also be problematic. Glass *et al.* (1981) acknowledge this problem and advise that the control group standard deviation is often the best choice, especially where more than one treatment group is compared with the same control (p107). Hedges and Olkin (1985) devote somewhat more space to the discussion of this issue (pp78-80) and show that where the assumption of equal population variances is reasonable, the use of a pooled estimate of standard deviation has smaller bias and variance, i.e. is a better estimator.

They also derive a 'correction factor' for the bias in using this estimate of effect size (p81). Hedges and Olkin give a formula (equation 15, p86) for the estimated variance of this bias-corrected estimate of effect size based on pooled standard deviation, and hence derive its standard error.

Given the arguments about whether 'absolute change' or 'residual gain' would provide a more appropriate measure of the effect of the intervention, and equally, whether the assumption of equal variances could legitimately be made, it was decided to calculate values for the 'effect size' using all four combinations and see if they were in fact different. Table 20 shows effect sizes calculated using both control group and pooled estimates of standard deviation (and standard errors for the latter) based on 'absolute change', i.e. the simple difference between attitude scale measures on the final (post-test) and initial (pre-test) questionnaires. Table 21 shows the same for residual gains.[15] It should be noted that the standard error for the effect size was once again calculated on the assumption that the values in each group were independent, and is therefore likely to be an underestimate.

Table 20: Effect sizes defined by 'absolute change' in attitude constructs

		Self-efficacy			**Achievement Orientation**			**Locus of Control**		
		mean	n	SD	mean	n	SD	mean	n	SD
Pre-test (initial)	Feedback	3.40	22	0.59	3.23	22	0.54	3.30	22	0.80
	Control	3.42	22	0.55	3.27	22	0.66	3.14	22	0.79
Post-test (final)	Feedback	3.41	22	0.66	3.32	19	0.86	3.22	19	0.76
	Control	3.51	22	0.43	3.39	21	0.53	3.35	21	0.62
		SD est from:			SD est from:			SD est from:		
Absolute change		ctrl gp	pooled		ctrl gp	pooled		ctrl gp	pooled	
Effect size (ES)		-0.18	-0.14		-0.06	-0.05		-0.45	-0.41	
Unbiased est of ES			-0.14			-0.05			-0.40	
Std error of ES			0.32			0.32			0.32	

[15] In calculating the effect size using residual gains, the difference between treatment and control groups should still be standardised against the standard deviation of the raw (post-test) scores, whether for control group or pooled (Glass *et al.*, 1981, p118).

		Attitude to ALIS			Feedback Anxiety			Self Confidence		
		mean	n	SD	mean	n	SD	mean	n	SD
Pre-test (initial)	Feedback	3.21	22	0.44	3.05	22	0.69	4.00	22	0.62
	Control	3.24	22	0.68	2.89	22	0.77	4.00	22	0.60
Post-test (final)	Feedback	3.20	18	0.75	3.02	18	0.97	4.08	19	0.49
	Control	3.58	21	0.43	2.82	21	0.73	3.90	21	0.59
		SD est from:			SD est from:			SD est from:		
Absolute change		ctrl gp	pooled		ctrl gp	pooled		ctrl gp	pooled	
Effect size (ES)		-0.81	-0.58		0.05	0.04		0.30	0.32	
Unbiased est of ES			-0.57			0.04			0.31	
Std error of ES			0.33			0.32			0.32	

		Feedback Desire			ALIS Value			ALIS Fairness		
		mean	n	SD	mean	n	SD	mean	n	SD
Pre-test (initial)	Feedback	3.74	22	0.58	2.93	22	0.74	3.07	22	0.79
	Control	3.57	22	0.78	3.14	21	0.73	3.17	22	0.91
Post-test (final)	Feedback	3.81	19	0.72	3.19	18	0.69	2.96	18	0.89
	Control	3.90	21	0.45	3.55	21	0.75	3.46	21	0.64
		SD est from:			SD est from:			SD est from:		
Absolute change		ctrl gp	pooled		ctrl gp	pooled		ctrl gp	pooled	
Effect size (ES)		-0.59	-0.45		-0.19	-0.20		-0.61	-0.51	
Unbiased est of ES			-0.44			-0.19			-0.50	
Std error of ES			0.32			0.32			0.33	

Table 21: Effect sizes defined by 'residual gain' in attitude constructs

	Self-efficacy			Achievement Orientation			Locus of Control		
Residual Gain	mean	n	SD	mean	n	SD	mean	n	SD
Feedback	-0.02	19	0.43	-0.03	19	0.54	-0.14	19	0.46
Control	0.01	21	0.39	0.02	21	0.36	0.13	21	0.39
	SD est from:			SD est from:			SD est from:		
	ctrl gp	pooled		ctrl gp	pooled		ctrl gp	pooled	
Effect size (ES)	-0.07	-0.06		-0.09	-0.07		-0.43	-0.39	
Unbiased est of ES		-0.05			-0.07			-0.38	
Std error of ES		0.32			0.32			0.32	

	Attitude to ALIS			Feedback Anxiety			Self Confidence		
Residual Gain	mean	n	SD	mean	n	SD	mean	n	SD
Feedback	-0.18	18	0.57	0.00	18	0.68	0.13	19	0.35
Control	0.15	21	0.37	0.00	21	0.64	-0.11	21	0.38
	SD est from:			SD est from:			SD est from:		
	ctrl gp	pooled		ctrl gp	pooled		ctrl gp	pooled	
Effect size (ES)	-0.76	-0.55		0.01	0.01		0.41	0.44	
Unbiased est of ES		-0.54			0.01			0.43	
Std error of ES		0.33			0.32			0.32	

	Feedback Desire			ALIS Value			ALIS Fairness		
Residual Gain	mean	n	SD	mean	n	SD	mean	n	SD
Feedback	-0.12	19	0.60	-0.10	18	0.64	-0.22	18	0.52
Control	0.11	21	0.32	0.09	20	0.57	0.19	21	0.48
	SD est from:			SD est from:			SD est from:		
	ctrl gp	pooled		ctrl gp	pooled		ctrl gp	pooled	
Effect size (ES)	-0.49	-0.37		-0.25	-0.26		-0.64	-0.53	
Unbiased est of ES		-0.37			-0.25			-0.52	
Std error of ES		0.32			0.33			0.33	

It can be seen that the effect sizes calculated from 'absolute change' (Table 20) are very similar to those calculated from 'residual gains' (Table 21). Only one of the attitude constructs ('Self Confidence') differs by more than 0.1 in both effect size estimates; 'Self Efficacy' differs by a similar amount on only the 'control group SD' estimate. The effect sizes calculated from pooled estimates of standard deviation (SD) generally agree well with those derived from control group SD. For two of the constructs, however, the two differ by more than 0.1 ('ALIS Fairness' and 'Feedback Desire') and one by more than 0.2 ('Attitude to ALIS') in both tables. For all these three, the effect size calculated using control group SD is bigger (in absolute terms) than that derived from a pooled estimate.

None of the 'pooled' effect sizes is larger than 1.96 standard errors (i.e. large enough to be statistically significant on a non-directional test at the 5% level), although the two largest effects ('Attitude to ALIS' and 'ALIS Fairness') both get larger effect sizes when the control group SD is used. Although some of the effect sizes are large enough to be of interest if replicated, the small size of the sample makes it seem plausible that they might not be replicated with a different sample. Conventional statistical wisdom would conclude that the effect of the feedback on teachers' attitudes was 'not significant'. However, the data are quite consistent with the inference that the feedback did indeed have quite a substantial effect on some of the attitudes. Only a replication of the study, though, preferably with a somewhat larger sample, could establish with any degree of confidence whether or not providing the kind of feedback given could be expected to alter attitudes.

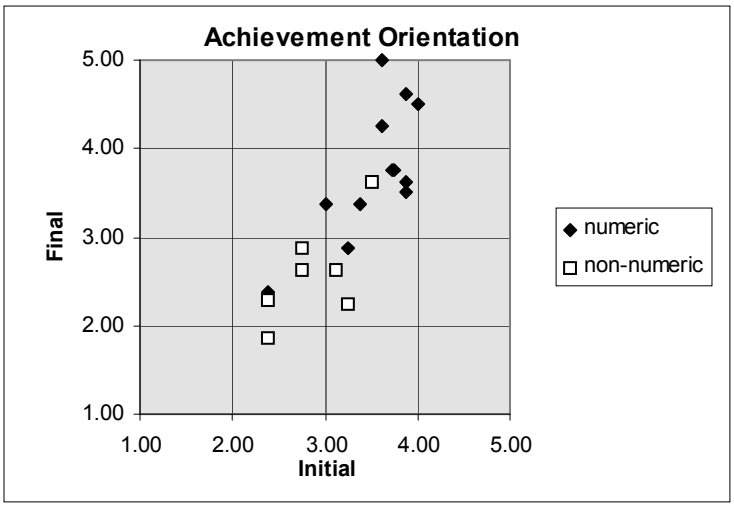
If the feedback was responsible for the differences found it is somewhat disappointing that the directions of change are often opposite to what might have been hoped. The attitude changes with the largest effect sizes ('Attitude to ALIS' and

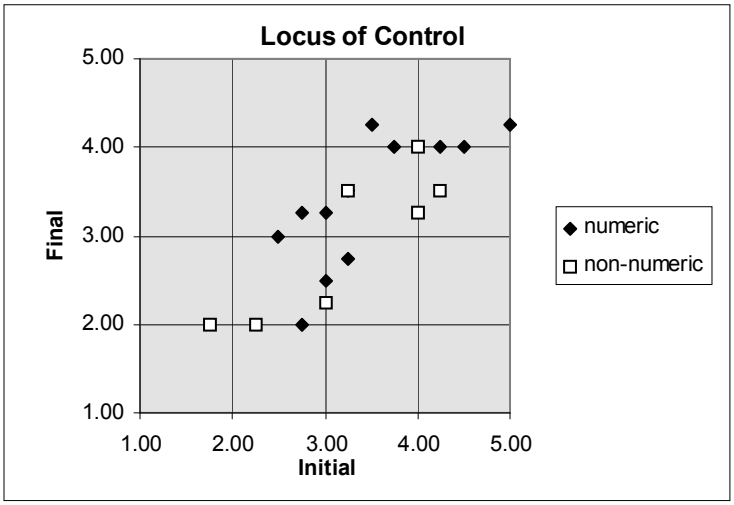
'ALIS Fairness') are both negative, suggesting that receiving feedback might have made people less positive towards ALIS and perceive it as less fair than those who received no additional feedback. Of course, it may be that too much or too incomprehensible feedback would result in more negative attitudes. In trying to account for these apparent attitude changes, therefore, it seemed likely that there might be interactions between a person's change in attitude and other variables such as how easy they had found the feedback to understand, their subject type or whether the content of the feedback had been positive or negative (i.e. the performance of their teaching groups). These interactions were therefore examined.

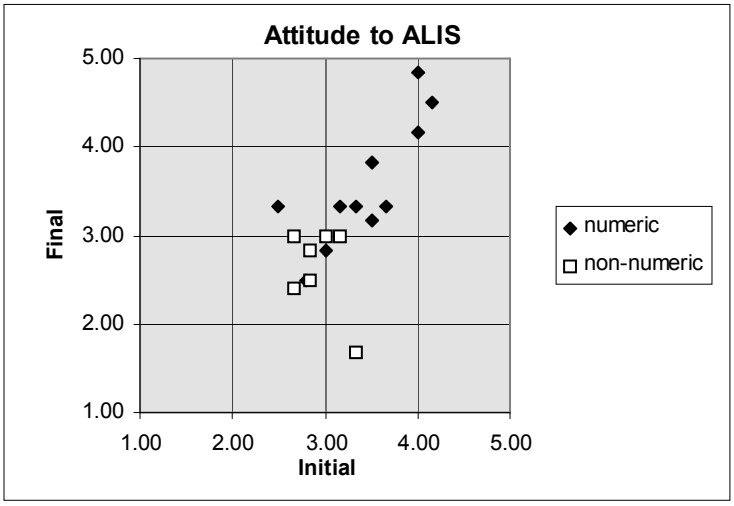
Differences between attitude changes by subject type

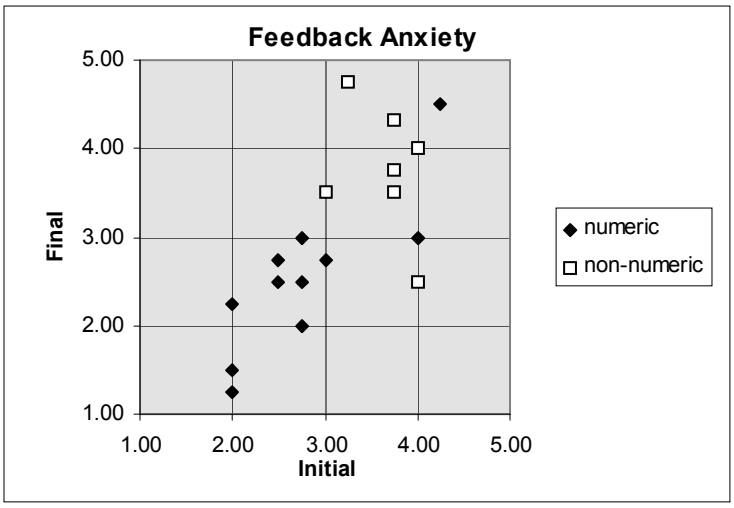
The scatter graphs showing final and initial attitudes (Figure 6) were redrawn to show only the feedback group, with the two subject types separated. These are shown in Figure 8. These graphs are one way to represent the data without relying on any statistics, and also avoid entering the debate about what kind of change scores to use.

Figure 8: Scatter graphs of initial and final scores on constructs, separated by subject type

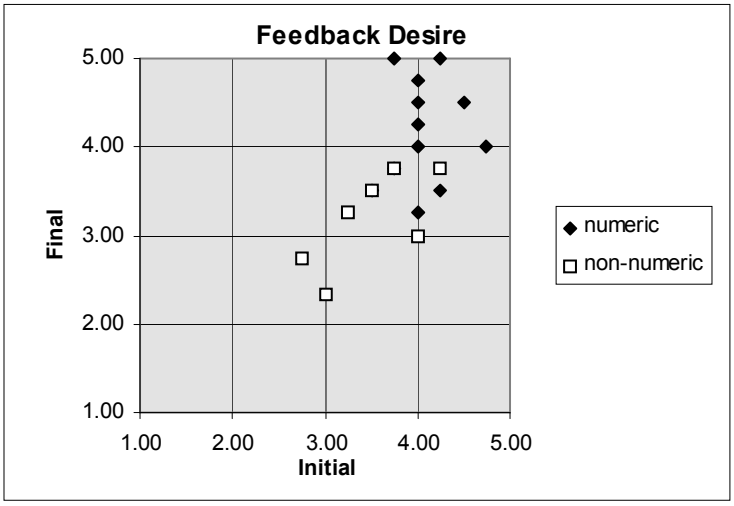

Achievement Orientation
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
numeric
non-numeric

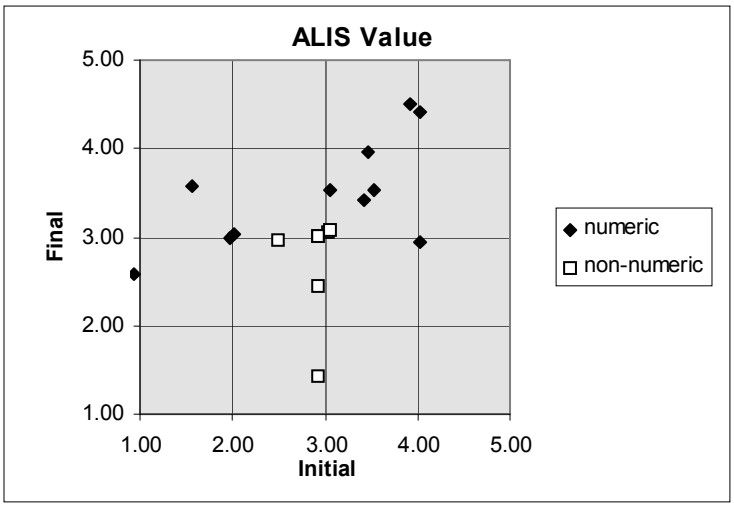

Locus of Control
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
numeric
non-numeric

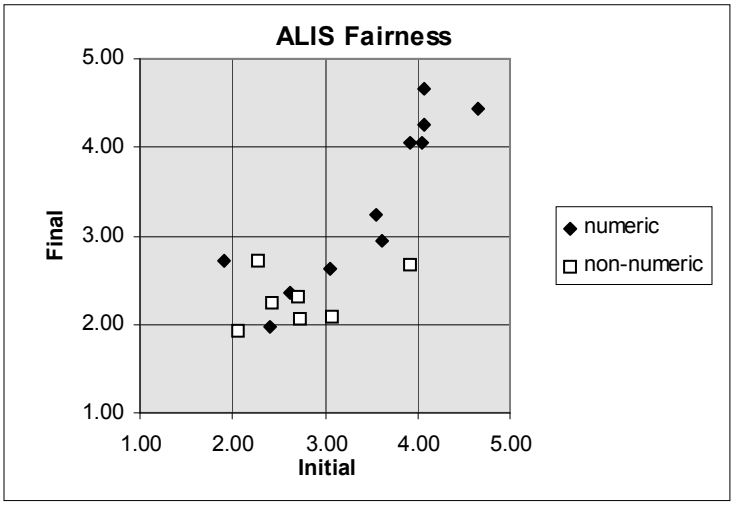

Attitude to ALIS
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
numeric
non-numeric


Feedback Anxiety
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
numeric
non-numeric


Self Confidence
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
numeric
non-numeric


Feedback Desire
5.00
4.00
3.00
2.00
1.00
Final
1.00
2.00
3.00
4.00
5.00
Initial
numeric
non-numeric





Immediately obvious from a number of these graphs is that there were substantial differences between the two groups before they received any feedback. In particular, the non-numeric teachers seem to have started with lower perceptions of their own effectiveness (self efficacy and self confidence), less inclination to receive feedback (lower achievement orientation, feedback desire and higher feedback anxiety) and a generally less positive attitude towards ALIS (lower scores on attitude to ALIS and ALIS Value). Mean values for the attitude scales for both subject types are shown in Table 22 for both feedback and control groups. Of course, it is impossible to say whether these differences would be found in the wider population of teachers, since the sample is so small and, as has already been said, not representative. However, it is interesting to note that the initial differences between the subjects for the control group are all in the same directions (but mostly smaller). It does seem plausible that teachers who are numerically minded might be more positive about

ALIS than others, since a good deal of the information ALIS provides is numerical or graphical. However, it is hard to see why they might have higher perceptions of self efficacy or be more disposed to seek feedback on their performance.

Many of these initial differences in attitude appear to have increased after receiving the feedback. The graphs of Achievement Orientation and Feedback Desire suggest that it is substantial increases for a small number of people in the numeric group that largely account for the widening of the gap. In fact, the same four people (the four with highest final scores) have the largest increases in both attitudes. The size of this subgroup of four is enough to suggest that they are not just outliers, but there may be some reason why their attitudes changed more, but they do not appear to be different from the others on any of the other variables collected. One of the four was subsequently interviewed.[16]

The changes on all three of the constructs which were measuring attitudes towards ALIS (Attitude to ALIS, ALIS Value and ALIS Fairness) suggest that the differences between the numeric and non-numeric teachers widened substantially after receiving feedback, with the numeric teachers becoming relatively more positive. This is evident from the means in Table 22. However, when the effect sizes for the feedback are calculated separately for the two subject types (Table 23) it is clear that the widening is mainly a result of the non-numeric group becoming more negative, while the numeric teachers' attitudes do not appear to have been greatly changed by the feedback. The graphs of Attitude to ALIS and ALIS Value show that the widening of the difference in means after receiving feedback is partly due to a single outlier, the point with the lowest final score on each. Again, this is the same person in both, a person whose attitude changes, as measured by the questionnaires, had apparently been uncommonly large on many of the constructs. He was subsequently interviewed,[17] and his comments threw considerable doubt on the validity of the questionnaire attitude constructs (at least for him).

The constructs Self Efficacy and Self Confidence appear from the graphs to have had significant initial differences between the subjects, but these seem to have

---

[16] This person was identified as 'A'. See Appendix 6K for a full transcript of the interviews, and Section 6.6 (p173) for analysis and discussion of them.

[17] This person was identified as 'B' in Appendix 6K.

been largely preserved by the feedback, rather than increased. This impression is confirmed by the mean values.

Table 22: Mean attitudes on initial and final questionnaire, separated by subject type

CONSTRUCT		MEAN ATTITUDE SCORE			
		Numeric subjects		Non-numeric subjects	
		Feedback (n=14)	Control (n=11)	Feedback (n=8)	Control (n=11)
Self Efficacy					
	Initial	3.64	3.51	2.98	3.34
	Final	3.77	3.57	2.80	3.44
Achievement Orientation					
	Initial	3.41	3.42	2.92	3.11
	Final	3.75	3.45	2.59	3.33
Locus of Control					
	Initial	3.43	3.02	3.06	3.25
	Final	3.40	3.23	2.93	3.48
Attitude to ALIS					
	Initial	3.34	3.27	2.98	3.21
	Final	3.56	3.53	2.63	3.63
Feedback Anxiety					
	Initial	2.77	2.75	3.53	3.02
	Final	2.55	2.75	3.76	2.90
Self Confidence					
	Initial	4.18	4.09	3.69	3.91
	Final	4.25	4.00	3.79	3.80
Feedback Desire					
	Initial	3.86	3.77	3.53	3.36
	Final	4.17	4.00	3.19	3.80
ALIS Value					
	Initial	2.93	2.95	2.94	3.35
	Final	3.50	3.27	2.71	3.85
ALIS Fairness					
	Initial	3.25	3.30	2.75	3.05
	Final	3.39	3.45	2.29	3.47

Note: Sample sizes quoted relate to initial questionnaire respondents. Not all questions were answered and four final questionnaires were not returned. Minimum numbers for each column are 11, 11, 7, 10 respectively.

Table 23: Effect size estimates for feedback effect on attitudes, separated by subject type

CONSTRUCT	NUMERIC				NON-NUMERIC			
	Absolute Change		Residual Gain		Absolute Change		Residual Gain	
	effect size	std error	effect size	std error	effect size	std error	effect size	std error
Self Efficacy	0.21	0.42	0.27	0.42	-0.33	0.50	-0.69	0.51
Achievement Orientation	0.27	0.42	0.28	0.42	-0.75	0.51	-0.81	0.51
Locus of Control	-0.35	0.42	-0.16	0.42	-0.69	0.51	-0.72	0.51
Attitude to ALIS	0.17	0.42	-0.12	0.43	-1.15	0.53	-1.42	0.55
Feedback Anxiety	0.07	0.42	-0.26	0.43	0.22	0.49	0.47	0.50
Self Confidence	0.44	0.42	0.46	0.42	0.55	0.50	0.35	0.50
Feedback Desire	-0.22	0.42	0.03	0.42	-1.25	0.54	-1.20	0.53
ALIS Value	0.42	0.43	0.39	0.43	-0.77	0.52	-1.06	0.54
ALIS Fairness	-0.23	0.43	-0.19	0.43	-1.14	0.53	-1.38	0.55

Note: Sample effect sizes are the difference between mean for feedback group and mean for control, standardised by pooled estimate of final questionnaire standard deviation, and corrected for bias. 95% confidence intervals for true effect size are given by [effect size – 1.96std error, effect size + 1.96std error].

Some of the effect sizes in Table 23 are fairly substantial, suggesting the feedback may have had quite an effect on attitudes for some people. For the numeric group, however, none of the effect sizes are large enough to be statistically significantly different from zero.[18] Nevertheless, effect sizes of the order of 0.4 (e.g. for Self Confidence, ALIS Value) would generally be considered quite significant, if replicated. For the non-numeric group some of the effects are really very large. Three of the attitudes (Attitude to ALIS, Feedback Desire and ALIS Fairness) have effect sizes with absolute value greater than 1, measured by both residual gain and absolute change. All six of these effect sizes are large enough to be considered statistically significant on a non-directional test at the 5% level. It seems reasonable to infer, therefore, that receiving the feedback caused the teachers of non-numeric subjects to have a generally less positive attitude towards ALIS, to be less keen to receive feedback and to perceive ALIS feedback as less fair than they had done previously.

[18] The standard errors quoted in Table 23 are once again calculated on the assumption that individual teachers' attitude scores are independent of each other, that is to say that knowledge of one teacher's attitudes should make no difference to the expectation about the attitudes of any other. However, some of the teachers in the same group were in the same department and it seems likely that some of their attitudes (e.g. towards ALIS) might be related. The effect of this clustering is to make the calculated standard error an underestimate of the true variability of the calculated effect size.

However, there were initial differences between the two groups in terms of their attitudes (see Figure 8, p157) and their students' examination performances (see Table 30 and Table 31, p187). It is therefore impossible to say whether their different responses to the feedback were a result of their subject type or because of other differences. Only a replication of the experiment in which treatment and control groups were better matched could establish this.

One further construct, ALIS Value, reached statistical significance and absolute value greater than 1 when its effect size was calculated from residual gain, but not when absolute change values were used, the latter reaching a still appreciable -0.77. The inference that the feedback caused the non-numeric teachers to perceive ALIS as having less value may therefore be less secure. Substantial (but not statistically significant) negative effects were also found for Achievement Orientation and Locus of Control. In both cases, non-numeric teachers' scores were reduced after receiving the feedback, though whether this was a causal effect or an accident of sampling is hard to say.

The question of the size of the differences between the two subject types can also be considered. The difference between the effect size estimates for numeric and non-numeric teachers is greater than 1 for four of the attitudes (Achievement Orientation, Attitude to ALIS, Feedback Desire and ALIS Value), whether measured by absolute change or residual gain. The 'residual gain' difference for ALIS Fairness is also greater than 1, and when measured by absolute change is close to this level. With the exception of Self Confidence (for which the subject difference is small and in opposite directions on the two methods), the other attitudes (i.e. Self Efficacy, Locus of Control and Feedback Anxiety) all show appreciable but smaller differences. It would thus be fair to say, for this sample at least, that there were substantial differences between the two subject types in the way the teachers responded to the feedback. In all cases (apart from Self Confidence) the feedback effects are more 'negative' for the non-numeric group than the numeric.

However, the small size of the sample once again means that the confidence interval around any estimate of the size of the differences is quite considerable. If the two estimates of effect size (i.e. for numeric and non-numeric) are assumed to be independent, then the variance of the difference between them will be the sum of the two variances. By this calculation, only one of the 18 differences (9 constructs, each

calculated by both residual gain and absolute change) is statistically significantly different from zero at the 5% level, a result entirely compatible with chance variation.

To summarise, although there are substantial differences between the effects of feedback on numeric and non-numeric teachers in the sample, statistical orthodoxy does not allow us to reject the hypothesis that the teachers could have been sampled from a population in which the two subjects were the same. We are, however, allowed to infer that, for some of the attitudes, the feedback did have a negative effect on the non-numeric group.

A number of warnings about the danger of over interpreting the data must be made here. Because the sample was so small (only seven non-numeric teachers received the feedback and returned the final questionnaire) and drawn from such a limited number of institutions, it is impossible to generalise the findings to any wider population. The results are at best suggestive, and their external validity can only be firmly established by replication. However, it is arguable that the results of any single experiment, no matter how large or what level of significance is reached, can only ever be suggestive: sound inference follows only from replication.

Furthermore, because of the initial attitude differences between the two subjects within the group who received the feedback (differences which were generally somewhat smaller in the control group), the control group may not have been well matched, and the inference that the differences in attitude change were caused by the feedback may be suspect. To put it more simply, if the non-numeric teachers in the feedback group did not start out with the same attitudes as those in the control group, it is hard to be sure that differences in attitude change were really caused by the feedback and would not have happened anyway, given their different starting points. Although the non-numeric departments were matched as well as possible before being randomly assigned to either feedback or control, the numbers were so small that this cannot really be considered an adequate guarantee of equivalence. This may be seen as a threat to internal validity, the attribution of the cause of the effects seen to the treatment differences.

Despite these reservations, the analysis of separate effect sizes for the two subject types seems to shed some light on the apparently anomalous results of Table 20 and Table 21, the effect sizes taken for the group as a whole. What appeared to be negative effects of the feedback are now seemingly limited to teachers of non-numeric

subjects. Given that these teachers generally reported that they found the feedback quite difficult to understand, it may not be too surprising if it often had a negative effect on their attitudes.

Interactions between attitude changes and other variables

For the group of teachers who received the feedback, correlations were calculated between their attitude changes and their perceptions of the feedback (Table 24), and the content of the feedback (Table 25). At this point it was decided to define attitude change in terms of absolute change since this would provide an unbiased, readily interpretable measure of change which is not subject to the variability of a regression equation derived from a small sample. Although the effect sizes calculated using residual gains have often been slightly larger than those derived from absolute change, these are less replicable, since a different sample would give different regression equations (perhaps quite significantly different, given the size of this sample) and therefore residual gain changes would have to be interpreted differently. On the other hand, provided the same test was used, the interpretation of absolute changes in attitudes would be unaltered. Moreover, the previous analyses using both absolute change and residual gains had seldom found much disagreement. In fact, the equivalent correlations to those in Table 24 and Table 25 were also calculated for residual gains and again similar results found.

Table 24: Correlations between attitude changes and perceptions of the feedback

ABSOLUTE CHANGE IN:	PERCEPTION OF FEEDBACK								
	Ease of understanding			Time spent			Usefulness rating		
	r	n	95% C.I.	r	n	95% C.I.	r	n	95% C.I.
Self Efficacy	0.03	13	[-0.47, 0.52]	0.35	14	[-0.16, 0.71]	-0.20	13	[-0.63, 0.33]
Achievement Orientation	0.28	13	[-0.25, 0.68]	0.07	14	[-0.42, 0.53]	0.28	13	[-0.25, 0.68]
Locus of Control	0.10	13	[-0.42, 0.57]	0.59	14	[0.15, 0.83]	0.28	13	[-0.25, 0.68]
Attitude to ALIS	0.45	13	[-0.06, 0.77]	0.23	14	[-0.29, 0.64]	0.41	13	[-0.10, 0.75]
Feedback Anxiety	0.02	13	[-0.48, 0.51]	0.28	14	[-0.23, 0.67]	0.15	13	[-0.37, 0.60]
Self Confidence	0.08	13	[-0.44, 0.55]	0.32	14	[-0.19, 0.70]	-0.15	13	[-0.60, 0.37]
Feedback Desire	0.11	13	[-0.40, 0.58]	0.37	14	[-0.14, 0.72]	0.30	13	[-0.23, 0.69]
ALIS Value	0.46	13	[-0.04, 0.78]	-0.33	13	[-0.71, 0.20]	0.48	13	[-0.03, 0.79]
ALIS Fairness	**0.52**	13	[0.03, 0.81]	0.34	13	[-0.18, 0.72]	0.47	13	[-0.03, 0.79]

Note: Correlations are calculated for feedback group only. 95% confidence intervals are derived from Fisher's Z-transform.

Given the apparent relationship between subject type and the effect of the feedback on some attitudes, it is perhaps surprising that these correlations are not higher. Although the differences in feedback effects for the two subject types were not large enough to dispel the explanation that this could have been simply a sampling phenomenon, for this sample at least, there were some appreciable differences. If it had been the difference in the ease with which they understood the feedback that made non-numeric and numeric teachers respond differently, then one might expect 'Ease of understanding' to be highly correlated with changes in attitude.

Of the four attitudes with the biggest subject difference in feedback effect (Achievement Orientation, Attitude to ALIS, Feedback Desire and ALIS Value) the correlations between attitude change and reported ease of understanding were 0.28, 0.45, 0.11 and 0.46, respectively. The attitude with the next largest subject difference (ALIS Fairness) has the highest correlation at 0.52. None of these are large enough to suggest that ease of understanding would be a good predictor of attitude change. A

correlation of 0.45, for example, indicates that just 20% of the variance in one variable is accounted for by the other; for 0.52 the figure is 27%. However, the measured correlation will be reduced by any measurement error in either variable, and also by the fact that 'Ease of understanding' was measured on a five-point (rather than continuous) scale and was some way from being normally distributed. The correction for attenuation can again be applied here (see p130). For example, for two variables with reliabilities of 0.7, an estimate of the 'true' correlation between the underlying constructs would be 0.64 if the measured correlation was 0.45.

The correlations between attitude changes and the other two variables, time spent and perceived usefulness, were of the same order or lower. Once again, it must be said that if these sample correlations are to be used as estimates for a wider population, the confidence intervals are so wide as to make them almost meaningless, even if the sample had been randomly selected.

Table 25: Correlations between attitude changes and content of the feedback

ABSOLUTE CHANGE IN:	CONTENT OF FEEDBACK								
	Students' mean Standardised Residual			Students' mean Relative Value Added			Students' mean Attitude to Subject		
	r	n	95% C.I.	r	n	95% C.I.	r	n	95% C.I.
Self Efficacy	0.04	19	[-0.39, 0.45]	0.13	19	[-0.31, 0.52]	0.30	18	[-0.15, 0.65]
Achievement Orientation	0.15	19	[-0.29, 0.53]	0.24	19	[-0.20, 0.60]	**0.51**	18	[0.10, 0.77]
Locus of Control	-0.02	19	[-0.44, 0.40]	0.00	19	[-0.42, 0.42]	0.24	18	[-0.21, 0.61]
Attitude to ALIS	0.10	19	[-0.33, 0.50]	0.21	19	[-0.24, 0.58]	0.36	18	[-0.08, 0.69]
Feedback Anxiety	-0.05	19	[-0.46, 0.38]	-0.08	19	[-0.49, 0.35]	0.28	18	[-0.17, 0.63]
Self Confidence	0.18	19	[-0.26, 0.56]	0.18	19	[-0.27, 0.56]	0.21	18	[-0.25, 0.59]
Feedback Desire	0.02	19	[-0.41, 0.44]	0.05	19	[-0.38, 0.46]	0.33	18	[-0.11, 0.67]
ALIS Value	0.39	18	[-0.05, 0.70]	**0.47**	18	[0.05, 0.75]	0.06	17	[-0.39, 0.49]
ALIS Fairness	0.11	18	[-0.34, 0.52]	0.25	18	[-0.20, 0.62]	0.09	17	[-0.36, 0.51]

Note: Correlations are calculated for feedback group only. 95% confidence intervals are derived from Fisher's Z-transform.

The correlation coefficients in Table 25 are, if anything, even less indicative of significant relationships. It might reasonably have been expected that those who received largely positive feedback might become more self confident, believe more in their own ability to influence their students' performance, become less anxious about receiving feedback and perceive the information and its source more positively. However, none of these changes were evidently associated with whether the feedback a person received was good or bad.

Possible explanations for this rather disappointing lack of associations include measurement error (as before) and the possibility that the teachers generally knew already how well their students had done, and thus changed little in response to feedback that told them nothing new. They would already have received the individual student residuals from ALIS and some had certainly already analysed them class by class. They might also have been able to compare their students' performance in their own subject with that in others (i.e. some measure of 'Relative Value Added' – RVA). Although they would not have had access to the students' 'Attitude to Subject' scores, after teaching them for up to two years they would no doubt have a fair idea of their attitudes.

Relationships between attitudes and past performance

Given that attitude changes did not appear to be related to the information about students' performance contained in the feedback, the question of whether there was any relationship between teachers' initial attitudes and the performance of their students was considered. In fact, attitudes measured by both the initial and final questionnaire were examined, and the average of the two. This latter seemed appropriate since the two were well correlated for most of the constructs so a measure which combined them might well be more reliable than either alone. For each of these three measures (initial, final and average) on each of the nine constructs, the correlation with students' performance (as measured by mean standardised residual and relative value added) was calculated.

At first it seemed that there were some significant correlations (of the order of 0.4 and large enough to reject the null hypothesis). However, when the scatter graphs

were plotted, it became obvious that the correlations depended heavily on two outliers: two teachers whose students' performance had been substantially better than anyone else's. These two had mean standardised residuals of 1.3 and 1.1, while all the rest were between -0.70 and 0.53. For such extreme values on one variable, a small change in the other variable for those two people would have a dramatic effect on the correlation coefficient. Although it was 'statistically significant', a statistic which was so dependent on the responses of just two people could not be considered very secure. When the two outliers were removed, most of the correlations were reduced, and no construct now had all three correlations above 0.2. Thus, it would be fair to say that teachers' attitudes, as measured by the questionnaires, were not significantly related to their students' performance.

Self-perception of changes

The second page of the final questionnaire asked respondents to describe any changes – in their attitudes towards ALIS, in how they would use ALIS feedback and in their teaching – that might have resulted from their involvement in the project. The comments written in response to these questions have been transcribed in Appendix 6J (p293). It had originally been intended to classify the comments 'blind' (i.e. without knowing whether or not they had received the feedback) into 'objective' categories and to analyse the results to see whether there were any differences between those who had been given the extra feedback and those who had not. However, the experience of trying to do this with the comments on the initial questionnaire (see p107) suggested that this might not be a very useful way to proceed. More helpful would be an interpretive approach in which each individual set of comments was seen as an expression of that person's perception of how they had changed.

A common response to the first request (to describe any 'changes in your attitude towards ALIS and the feedback it provides') was to answer 'none' or 'not much'. Of those who did describe changes, a number did seem to have become more positive, but these were as likely to have been in the control group as not. Of course it is hard even for the respondents themselves to identify the cause of any changes in their attitudes, but with one possible exception (comment 16 – and see below for further consideration of this one), none of the changes described are clearly

attributable to the feedback supplied in this experiment. Many comments referred to specific reservations about ALIS, for example the ambiguity of its 'Perceived Learning Activities' (comment 8) or the limitations of GCSE performance as a predictor of A level (comments 26, 33, 39, 40).

The descriptions of any 'changes in how you will use ALIS feedback in the future' were equally hard to attribute to the feedback sent. Again, many people said 'none' or 'unsure' and as many changes seemed to have been described by those who did not receive the feedback as by those who did. One comment (16) referred specifically to two aspects of the feedback which were not part of the information otherwise sent by ALIS: target grades and data for comparing students' performance with their other subjects. This respondent had referred to a 'deepen[ed] interest' in the former and a realisation of how useful the latter could be in his description of attitude changes, and now said 'I will think about using MPGs [minimum predicted grades] in review and targeting sessions with students. I will analyse comparative data (as HoD) more if it is available easily'. Some comments, however, suggested that although many of the teachers were changing their practice in a number of ways, these changes could not really be attributed to the extra feedback sent in the experiment. For example, several people in the control group referred to the use of target setting (comments 7, 8, 24, 27) and a comment by one of those who had received the feedback (39) that '[I] will still analyse the results for individual sets' suggested that another feature of the feedback that took it beyond what had been provided by ALIS (i.e. the set by set analysis) was in fact being practised already.

Finally, the descriptions of 'any changes in your teaching' were generally non-committal and, once again, failed to provide evidence of the perception of any effects of the experimental feedback on teaching behaviour.

View of who should receive feedback

Respondents were asked whether they thought class by class feedback should be provided by ALIS and whether it should be sent only to the individual teacher(s) concerned, to the head of department and/or to the ALIS coordinator in the institution. The responses for those in the feedback and control groups are shown in Table 26.

Table 26: Frequencies of opinions about who should be sent class by class feedback

Feedback should be sent only to individual teacher	yes	no	no opinion
Feedback	5	6	5
Control	8	5	4

Feedback should be sent to Head of Dept	yes	no	no opinion
Feedback	10	2	5
Control	14	1	3

Feedback should be sent to ALIS coordinator	yes	no	no opinion
Feedback	8	1	7
Control	9	2	8

The word 'only' in the first statement was included in order to stress that this option could preserve the confidentiality of sending the feedback only to the person involved. However, it was logically unnecessary, since respondents could say 'no' to the suggestion of sending it to the other two people, and more importantly, it ruled out the possibility that the feedback should be sent to both the teacher concerned and to someone else. In fact this last combination was the choice of the majority. Of the 13 people who thought class by class feedback should be sent to the class teacher, 8 also thought it should go to the head of department (of whom 3 crossed out the word 'only' on the questionnaire) and the remainder either ticked 'no opinion' or left blank the questions about whether it should go to the other two. Thus nobody appeared to think that it should go to the class teacher and no one else.

In none of the three parts of Table 26 is there a clear difference between those who received the extra feedback and those who did not. The frequencies were also examined for differences between the two subject types and for different positions of responsibility within the institution. None of the differences were significant, either in terms of the apparent size of the difference or by the result of a chi-squared test for independence at the 5% level.

The overall opinions expressed are nevertheless quite interesting. While views about whether the individual teacher should receive class by class feedback were more

or less equally divided (13 for, 11 against), there was an overwhelming majority in favour of it being sent to the head of department (24 for, 3 against). Almost as large was the majority for sending it to the ALIS coordinator (17 for, 3 against). These results are somewhat at odds with the assumptions made at the outset of the project that teachers would generally like to have good feedback about their own performance, but that they might feel some anxiety about the wider dissemination of any analysis that could be used to judge them. It must of course be remembered that these results came from institutions that had been using ALIS data for several years and may not be typical of other schools and colleges.

6.6 FINAL INTERVIEWS

The objectives for the interviews were largely twofold: to validate the inferences from the questionnaires and to provide greater insights into the teachers' uses and perceptions of the feedback. Five questions were asked which were similar in content to some of those on the questionnaires. These concerned respondents' locus of control, their self confidence about their teaching effectiveness, their perception of the fairness of the ALIS feedback, their general attitude to ALIS, and the ease with which they had understood the additional feedback (if they had received it). The relevant questionnaire responses of the six people interviewed are shown in Table 27.

Table 27: Questionnaire responses of interviewees

		INTERVIEWEE						WHOLE SAMPLE (n=44)	
		'A'	'B'	'C'	'D'	'E'	'F'	Mean	S.D.
Feedback/Control		Fbk	Fbk	Ctrl	Fbk	Fbk	Fbk		
Ease of understanding		Very easy			Easy	Easy	Impossible		
Locus of control	Initial	5.00	4.25	3.25	3.75	3.25	3.00	3.22	0.81
	Final	4.25	1.75	3.25	4.00	3.50	2.25	3.29	0.70
Self confidence	Initial	5.00	3.50	4.00	4.00	4.00	3.00	4.00	0.62
	Final	5.00	4.50	4.50	4.00	4.00	3.00	3.99	0.56
ALIS Fairness	Initial	4.67	3.00	3.67	4.00	4.00	2.67	3.12	0.86
	Final	4.33	2.00	4.67	4.67	4.33	2.33	3.23	0.81
Attitude to ALIS	Initial	3.50	3.33	3.50	4.00	4.17	3.00	3.23	0.58
	Final	3.83	1.67	4.17	4.83	4.50	3.00	3.40	0.64

It is quite difficult, however, to extract from the interview data anything to compare with the questionnaire responses. Even when the person was able to quantify their attitude, it is not clear what would constitute agreement between the two measurements. For example, in the first interview, 'A' seemed happy to rate each attitude on a scale from 0 to 10. His ratings are all in the same direction from the sample mean as his questionnaire responses, and with the possible exception of 'Attitude to ALIS' (where he rated his attitude as 9, but his questionnaire scores are not much above the group mean), the sizes of the ratings seem appropriate for the scores on the relevant questionnaire constructs. However, this judgement of appropriateness does seem a bit arbitrary.

The second interview was with 'B', a person who was chosen on account of his apparently large changes in attitude between the two questionnaires. It became evident in this interview that these were questions to which he could not happily give a simple numerical answer, and this is probably a large part of the explanation of why his attitudes had appeared to be so erratic. In reply to the first question, about the extent of his control over student performance, his concern was to be consistent with his previous response – 'I think I must have said 5' – and it seems significant that his choice was for the neutral value, 5. He then went on to explain the difficulty of

summarising his attitude in a single number, finishing with the statement: '… the answer is it's so complex that I don't think one could say one's pupils as a whole …'

The projection of a complex attitude onto a numerical scale was, at least for him, not a meaningful activity. As the interview progressed and he was increasingly given permission to reject the '0 to 10' scale, it became clearer that he preferred to answer with the subtlety of words than the 'precision' of numbers. Although it was the interviewer who offered that, 'I think what you're saying really is that you can't translate it into a number isn't it?', his reply 'Yes, exactly' indicated clear agreement.

It was perhaps unfortunate that this person was the subject of the second interview. I was quite sympathetic to this reluctance to quantify a complicated issue and took 'B's attitude as indicative of a more general danger of over-simplification by quantification. Consequently the '0 to 10' scale was abandoned in the subsequent interviews; had it not been for the 'extreme' case of 'B', I might well have persisted with it. However, the complexity of the answers given by the other people interviewed makes it hard to see how they could meaningfully be translated into a single number. For example, in response to the same question about the degree of control over student success or failure, 'E' said:

> Well, sometimes I really think I've helped out students a lot and made a difference, and other times I think that no matter what I'd done the student would have got an 'A' anyway, or would have failed anyway. … I think the teacher can make a big difference in some cases especially if the student is receptive to that. In other cases, the student's attitude makes it difficult for the teacher to make a big difference.

This illustrates clearly the limitations of a quantitative methodology in trying to understand or represent fairly anything as subtle as the attitudes and perceptions involved here.

Another issue that relates to the validity of the interpretation of the questionnaire constructs was the interviewees' perceptions of their attitude changes. When asked whether the difference in responses on the two questionnaires was significant, 'C' replied,

> I suspect that it might not have meant anything. … It depends how you were feeling at the time.

When questioned about her own increase in score on the 'ALIS Fairness' construct, 'C' said, 'Sounds pretty arbitrary to me.' Later she commented:

> I think I think different things, you know, which probably is the reason why I answered differently maybe the second time. I didn't look back on what I'd answered the first time…

The changeable nature of perceptions of self-confidence was also acknowledged by 'D':

> It depends on … well take today for instance – I'm feeling quite good. I had two really good lessons this morning. Tomorrow, I'll probably have an extremely bad one and feel awfully depressed and give you a different answer. It's sort of patchy.

Another difficulty with interpreting any apparent changes is that, when asked how their attitudes had changed, many of the replies were not really focused on the specific attitude, but on other issues that seemed to be more important to the person being interviewed. For example, in the interview with 'A', he rated his confidence in his effectiveness as a teacher as 'about 9 or 10', but qualified it by saying that he was, 'becoming incredibly disillusioned'. When asked whether his attitude had changed, his reply was clearly about his feeling of disillusion, rather than his confidence in his effectiveness. This shift of focus was typical of a number of replies to the question about how people's views had changed.

To summarise the contribution the interviews made to the validation of the questionnaire interpretations, therefore, it seems that the evidence from them was, at best, equivocal. The interviews arguably did more to undermine confidence in the validity of the questionnaire constructs, and in particular, the changes in them, than to endorse their previous interpretation.

On the second objective for the interviews – to throw light on people's perceptions of ALIS and the feedback – the results were more encouraging. Unfortunately, when people were asked specifically about the feedback sent as part of this project, they were generally unable to remember sufficiently clearly what it had been (for example, 'D': 'I can't quite remember what it was now'). In particular, the interviews were able to provide little insight into the difficulties of understanding that some teachers had had. This was an unfortunate consequence of the time delay in receiving all the final questionnaires, and the need not to conduct the interviews until they had all been returned. However, some comments made did refer specifically to the feedback sent in the experiment, while others referred to the feedback routinely sent by ALIS.

As in the questionnaire responses, the phenomenon of the numerical/non-numerical divide in ease of understanding was witnessed again for the wider ALIS feedback in a comment by 'D':

> The first two or three years we were using it, it was very difficult to get the people who weren't mathematically trained to actually understand what the information was. That's got a lot better and it's part of the culture here, so no one has any real worries about it, but there are certainly certain aspects of it which people find a bit scary perhaps. They're not really quite sure what they are looking at and they'll come along and ask me – ask the scientists – what it actually means. It's fair enough.

The two teachers of non-numerical subjects interviewed ('B' and 'F') were also the ones to express reservations about the complexity of the ALIS feedback. 'F' referred to being 'faced with sheets and sheets of statistics,' and agreed that the numbers were off-putting. Interestingly, although he rather modestly described himself as confident 'with the very elementary bits of it', 'B' went on to describe some quite sophisticated uses of the ALIS feedback, analysing the performance of individuals and of teaching groups. Nevertheless, he did say that '… my gut feeling is that it seems a bit more complex than I want it to be.'

Some of the interviewees seemed a little embarrassed to be asked to express an opinion about their own effectiveness, and this makes their comments even harder to interpret. 'D's response is a good example of this:

> But on the whole I do a good job. The students tell me I do a good job anyway. Perhaps it's me just being hypercritical of myself. Put it this way, I know I could do a better job – that's probably the best way of putting it. That's probably the key issue from my own point of view.

There were some interesting comments made on the question of teachers' perceptions of their responsibility for student achievement. 'A' accepted a large part of this responsibility (rating this at 'about 7 or 8') and attributed his students' considerable success to 'the huge effort we put into students here.' Comments by others suggested that they felt their responsibility was less with larger groups ('C') or with less able students ('F'). The complexity of this issue was widely acknowledged. 'B's description of the different outcomes of three different groups in which he had shared the teaching illustrates the difficulty of attributing responsibility:

> One group I took over from two members of staff who left, so both of their teachers left. I and a probationary teacher who had just joined the college staff

> took them over. They were the usual mixture of hard working and not very hard working – within the same person sometimes! And in the end their results were really pretty good. I had another set who were totally boring and their results were boringly predictably pretty good, and then I had another set who had always been not very good on attendance and several people left it and illness and all sorts of psychological traumas – real awful things – and their results on the whole I felt were a bit disappointing, even taking that into account. So that's me, the common link between those three sets, with a great variety of people I was sharing with too – experienced and probationary. When you throw in all those other things that ALIS looks at – social background and all that sort of stuff – you do end up wondering what it's telling you that's of any use.

However, even for an individual student it can be very difficult to assess the significance of the teacher's contribution. 'B' again:

> I should think that he'd missed a good quarter of the lessons… And he was obviously naturally quite bright. He had a lot of problems at home, stress and strain and blah, blah, blah, and I was thinking, 'Oh well, he'll probably get a D if he's lucky,' and lo and behold, he gets a B. He came in yesterday to say 'I came to say thank you' and I really felt 'What are you thanking me for? Apart from the fact that every now and again I chivvied you and when you were there I did my best and so on, but simply in terms of hours of contact it couldn't have made all that much difference.'

Attitudes to ALIS were also interesting. 'B' expressed the view that, although ALIS seemed to be a 'worthy enterprise' and 'there is something there that is a good idea', raw results were generally perceived as being more important than value added:

> … at the beginning of term you know when [the Principal] says, 'Well thanks everybody, great results…' … It doesn't strike home for us, I think, that we've got a good, or whatever, value added score because in society at large everybody is saying 'Winchester is wonderful because a hundred percent get grade A' and so on. … it seems slightly tangential to the main source of pleasure, which is, 'Oh great, 35 people got As and that was 20 per cent of the year and that's better than last year' – that sort of thing. We still seem to think of it in those terms.

However, he did concede that '… the big advantage of ALIS is that it's made everybody think about it.'

'A' clearly believed that the value added measures provided by ALIS were a fair measure of performance: 'I think I would put 10 at that. 9 or 10 anyway.' However, he described the student attitude feedback as 'boring' – although his comments suggest that he may have been thinking more of the data from questionnaires produced within his institution than of the ALIS feedback.

'C' described how her department had used ALIS feedback to help make decisions about which syllabuses were most suitable for their students, which suggests a fair degree of faith in its validity. However, she did feel that when one syllabus was attracting a disproportionate number of less able students this had an effect on the results, and the ALIS analysis was not able to take this into account. In fact a number of factors not taken into account by ALIS were seen as shortcomings by others. 'F', who when asked for her attitude to ALIS said, 'I don't really take much notice of it', listed her reservations about using it to make predictions:

> … it just doesn't take personality into account, and it doesn't take, you know, time constraints, pressures that come on them during the two years they're here.

Also mentioned, by 'D', was the effect of the previous school on the intake measure, and the need to take that into account in interpreting GCSE scores – something ALIS cannot do.

Other reservations expressed were more to do with the need for care in interpretation than any genuine failings of the value added model. For example, 'D':

> It's a useful little tool as long as you are conscious of its limitations. I think that's the danger. If people don't understand the system terribly well it can be used as a blunt weapon, lacking finesse.

Similarly, 'E':

> … as a measure for judging added on value I think it works quite well but I think one's got to be careful of applying it in individual circumstances.

A slightly more telling criticism was implied by 'B's observation that the first year's value added results had been quite positive, while subsequent years had seen performances close to average, despite his feeling that nothing significant had really changed in the quality of the teaching. His comment was that 'it's not reliable.' This suggests that he was interpreting the value added feedback as a measure of teaching effectiveness (which one would expect to be fairly constant) rather than as a measure of student progress (which would presumably be affected by a number of factors not measured by ALIS, and therefore vary appreciably).

Finally, it must be noted here that none of the interviews followed closely the interview schedule drawn up beforehand, and some of them may seem to bear very little relation to it. This was owing in part to the difficulty of concentrating

simultaneously on what was being said, the schedule and its objectives, and the tape recording process – and no doubt also in part to the inexperience of the interviewer. However, it was also to some extent a deliberate choice to allow the interviewees free expression and to follow up and respond to whatever they said. To what extent the resulting differences in content and structure should be seen as a threat to the reliability of the data gained, or whether the uniqueness of each adds to the validity of its interpretation is a matter for judgement (Hull, 1985).

6.7 EXAMINATION PERFORMANCE

Models used in analysis

It is now widely accepted in research on 'effectiveness' that one should 'pay attention to the multilevel organisational structure in which education occurs' (Raudenbush, 1989, p721), in other words, to use multilevel (ML) models to analyse the data.

The data in this study consisted of individual pupil-subject level outcome (A level grade, attitude to the subject) and intake variables (prior attainment at GCSE, sex, parental occupation). However, a number of pupils had data for more than one subject (e.g. they had taken both physics and mathematics) and performances by the same pupil in different subjects (even after controlling for intake) were found not to be independent. Pupils in each subject were nested within teaching sets, which were nested within examination years. However, teaching sets were also nested within 'teacher-combinations' and the same combination of teachers was often found teaching groups across more than one year. Teacher-combinations were nested within departments. Even this amount of complexity still makes no attempt to isolate the effect of an individual teacher, who may have taught some groups alone and others in combination with different colleagues. There was also the potential problem of a teacher who taught more than one subject, though fortunately (!) none of the data analysed in this study presented that particular complication.

Despite these complexities, it was possible to fit a number of multilevel models to these data and the advantages of ML modelling in taking account of the

relationships among variables at different levels – quite apart from its status as the orthodox methodology – required that it should be applied. The value added analysis routinely provided to schools by ALIS, however, uses residuals derived from ordinary least squares (OLS) regression of A level subject grade on average GCSE score. In order to make any feedback sent to be consistent with what might have been previously sent by ALIS, the performance analysis sent to teachers used these same residuals, both for individual pupils and averaged at the level of the teaching set. Three factors motivated the decision to continue with this model of value added alongside the multilevel model: firstly, the finding reported in the Value Added National Project (Trower and Vincent, 1995) that agreement between average residuals calculated by OLS and those derived from ML models is extremely good; secondly, the small size of the sample and consequent large standard errors in the estimation of the parameters in the ML model might well mean that residuals based on an OLS regression equation incorporating the whole of the A level entry in that subject in the ALIS project that year could be more reliable than the ML residuals derived from a small sub-sample; and, thirdly, that if the results of the study were to be fed back to the participants (some of whom had expressed considerable interest in its findings), it would be much better to be able to do so in terms with which they were somewhat familiar. Thus two analyses of student performance were conducted in parallel: one using a multilevel model, the other using OLS residuals.

Finally, an attempt was made to cut the Gordian knot of isolating some measure of individual teacher 'effectiveness' from data in which a large proportion of students were taught by more than one teacher. To do this, for each teacher the (value added) performance of each student taught by them was weighted by the proportion that teacher had contributed to the teaching of the group. For example, if a teacher had taught one group alone and another shared equally with another teacher, the results of the students in the second group were given half the weight in the calculation of that teacher's overall average. Clearly, this was a crude attempt to solve the problem, and, in particular, by treating the outcomes from a shared group as the sum of the individual teacher effects, it ignored any effect of the interaction between the two (or more) teachers. Ideally, a cross-classified multilevel model would have been used, but operational difficulties in getting the ML program to run this model successfully with the data and lack of time prevented this.

Implications of missing data

The problems of collecting the data, and the resulting gaps, have been described in Chapter 5. In terms of student numbers, 7% of the 'pre-test' measure (i.e. 111 out of the 1565 examination results from 1994-6) came from departments whose 1997 results were not available – a small but possibly significant proportion. These 111 results were nevertheless included in the analysis, since they were part of the experimental sample at the time of the random allocation to treatments and should therefore be included in any baseline measure. However, it must be remembered that, had the full sample been available in 1997, the figures for that year might have been different from the figures derived from the diminished sample.

When considered in terms of the number of teachers involved, the significance of the loss of data seems somewhat greater. Of the 44 teachers randomly allocated to either the feedback or control groups, only 31 (16 in the feedback group and 15 in the control) remained in the final dataset. The loss of almost 30% of the sample, in terms of the number of teachers involved, may be seen as a significant threat to making causal attributions for any differences found between the two groups. Owing to the scale of this sample attrition, the analyses which focused on the teacher as the unit (see below, p203) were restricted to the teachers for whom full data were available. This section may therefore be viewed as more of a quasi-experiment than a true experiment, since the equivalence between the two groups that the latter guarantees by random allocation was largely undermined by the loss of data.

Unadjusted characteristics

Table 28: Characteristics of feedback and control group students each year.

YEAR		A level grade (A=10, B=8, etc)		Avg GCSE score (A*=8, A=7, B=6,)		Parental occupation (1=unskilled, … 6=professional)	
		feedback	control	feedback	control	feedback	control
1994	mean	**5.42**	**5.92**	**5.82**	**6.07**	**4.77**	**4.62**
	n	221	203	221	203	221	203
	s.e.	0.21	0.23	0.05	0.05	0.06	0.23
1995	mean	**5.77**	**5.97**	**5.85**	**5.99**	**4.60**	**4.74**
	n	261	243	261	243	248	235
	s.e.	0.19	0.22	0.05	0.04	0.07	0.22
1996	mean	**6.01**	**5.89**	**6.14**	**6.10**	**4.72**	**4.73**
	n	353	284	353	284	321	259
	s.e.	0.17	0.19	0.04	0.05	0.06	0.20
1997	mean	**6.45**	**5.63**	**6.18**	**6.19**	**4.75**	**4.68**
	n	285	219	285	219	259	184
	s.e.	0.17	0.20	0.04	0.05	0.06	0.22

It can be seen from Table 28 that students in feedback and control groups were reasonably well matched in terms of incoming average GCSE scores and parental occupations in all four years. A level grades are also quite similar in 1994-6, with any differences between the groups always compatible with the differences in intake. In 1997, however, after the feedback was sent, students in the feedback group achieved just under half a grade (0.82 on the 'one grade = two points' scale) better than the control, with no corresponding difference in average GCSE scores or parental occupation. In terms of an effect size, this difference is equal to 0.28.[19]

When the changes in A level grade averages for both groups are calculated, relative to their pre-1997 averages, the effect of the feedback appears to increase slightly. The difference between the feedback and control groups is now almost exactly half a grade (0.96 on the points scale) and the effect size rises to 0.30.

[19] Unless stated otherwise, effect sizes have been calculated using a pooled estimate of standard deviation. Standard errors for these effect sizes are generally very small (approximately 0.01), since the groups now contain several hundred values. However, these standard errors still fail to incorporate the effects of clustering, and are therefore not very meaningful.

Residual gain analysis

Analysis of individual student performance and attitudes

The three performance measures which were included in the feedback sent to the teachers in the feedback group (Standardised Residuals, Relative Value Added and Attitude to Subject) were used to investigate any differences between the feedback and control groups. As explained in Chapter 1, individual students' Standardised Residuals are part of the feedback provided by ALIS to its schools. They are based on OLS regression of the A level grade (coded as A=10, B=8, C=6, etc) on average GCSE score (average of all grades achieved, coded as A*=8, A=7, B=6, C=5, etc), the regression equation being calculated separately for each A level subject. Residuals are standardised by dividing them by the residual standard deviation in order to make them comparable across subjects and years when the strength of the correlation (typically around 0.6) varies. The residual standard deviations for the data used here varied between 2.3 and 3.0. Differences in average Standardised Residuals could therefore only roughly be converted into equivalent differences in A level performance, but an approximation was achieved by taking 2.6 as an average value for the standard deviation and remembering that A level grade is measured on a scale where one grade equals two points. Thus, a difference in average Standardised Residuals of 1 was taken to be roughly equivalent to 1.3 A level grades. Average Standardised Residuals may therefore be interpreted as a measure of A level performance when the likely effects of prior achievement are (at least partially) allowed for.

'Relative Value Added' was invented for this project and was calculated by finding a student's average Standardised Residual in all their subjects and subtracting this from their Standardised Residual in the subject concerned. Its interpretation therefore depends largely on how one chooses to account for the correlation between the same student's Standardised Residuals in different subjects. If it is held that the correlation is a result of shared error (e.g. measurement error in the control variable or the effects of unmeasured variables such as general motivation or personality variables), then by comparing performance in one subject with that in others, one is, to some extent, adjusting for this error and thereby achieving a more valid measure of the

effect of teaching. If, on the other hand, the explanation for the correlation is held to be in the interaction among different subjects (a student whose work is good in one subject will find a pay-off in their other subjects) then subtracting the average 'value added' takes away some of the genuine achievement in the subject in question.

Attitude to Subject is a scale calculated from eight Likert-type items (reliability = 0.8) on the 'Extended ALIS' questionnaire. The scale goes from 1 (negative) to 5 (positive).

The means of each of these outcomes in each of the four years for which data were available are presented in Table 29, with results separated by treatment.

Table 29: Outcomes for students in feedback and control groups, 1994-7

YEAR		**Standardised Residual**		**Relative Value Added**		**Attitude to Subject**	
		feedback	control	feedback	control	feedback	control
1994	mean	**0.03**	**0.05**	**-0.11**	**-0.13**		
	n	221	203	212	200		
	s.e.	0.07	0.07	0.05	0.05		
1995	mean	**0.14**	**0.16**	**-0.07**	**-0.15**	**3.33**	**3.20**
	n	261	243	251	239	213	161
	s.e.	0.07	0.06	0.04	0.04	0.05	0.06
1996	mean	**0.03**	**0.04**	**-0.03**	**-0.05**	**3.45**	**3.52**
	n	353	284	351	278	143	130
	s.e.	0.06	0.06	0.03	0.04	0.06	0.06
1997	mean	**0.21**	**-0.06**	**0.02**	**-0.14**	**3.38**	**3.33**
	n	285	219	270	189	251	166
	s.e.	0.05	0.06	0.04	0.04	0.04	0.06

It can be seen from Table 29 that students' examination performance, whether measured by Standardised Residuals or by Relative Value Added, is very similar for the two groups in the years 1994, 1995 and 1996.[20] In 1997, however, both measures show a significant advantage to those in the feedback group. As mentioned above, it is impossible to translate a difference in Standardised Residuals precisely into A level

[20] An example of the effect of restricting the averages for 1996 to those students in departments which also had 1997 data can be seen in the figures 0.02 and 0.04 for Standardised Residuals and -0.04 and -0.06 for Relative Value Added for feedback and control groups respectively. From this it seems unlikely that the availability of their results in 1997 would have made a substantial difference.

grades, but the difference here is of the order of one third of a grade. To summarise, it would be fair to say that the teachers who were to receive the feedback were well matched with those who were not, in terms of their former students' A level grades, when adjusted for prior achievement. After the feedback was sent, students whose teachers received it achieved about a third of a grade higher (after adjusting for prior achievement) than those in the control group. The interpretation of the changes in Relative Value Added is similar. When adjusted performance in the 'experimental' subject is compared with that student's (adjusted) performance in their other subjects, students whose teachers had received the feedback outperformed those in the control by about a quarter of a grade.

Changes in students' attitudes are less clear. Unfortunately, no attitude data were available for 1994, so only two years' worth of data make up the 'baseline'. There appears to be a fair amount of variation in attitude scores, with the direction of the difference between feedback and control groups changing annually and being smallest in the year following the receipt of the feedback. Hence, there are no clear changes in Attitude to Subject.

As the teachers' attitude changes in response to the feedback appeared to have been related to their subject (see pp157-166), it was decided to split the sample into Numeric and Non-numeric sub-groups and repeat the above analysis.

Table 30: Outcomes for students in numeric subjects (Mathematics and Physics)

YEAR		STUDENTS IN NUMERIC SUBJECTS					
		Standardised Residual		**Relative Value Added**		**Attitude to Subject**	
		feedback	control	feedback	control	feedback	control
1994	mean	**0.35**	**-0.02**	**0.01**	**-0.18**		
	n	79	139	76	137		
	s.e.	0.12	0.08	0.07	0.05		
1995	mean	**0.40**	**0.05**	**0.08**	**-0.25**	**3.39**	**3.10**
	n	101	167	100	163	79	93
	s.e.	0.11	0.08	0.06	0.04	0.08	0.07
1996	mean	**0.19**	**0.06**	**0.09**	**-0.06**	**3.45**	**3.67**
	n	203	140	202	134	143	6
	s.e.	0.07	0.09	0.04	0.05	0.06	0.34
1997	mean	**0.27**	**-0.07**	**0.02**	**-0.12**	**3.37**	**3.11**
	n	148	119	139	91	127	78
	s.e.	0.07	0.09	0.04	0.05	0.06	0.07

Table 31: Outcomes for students in non-numeric subjects (English and French)

YEAR		STUDENTS IN NON-NUMERIC SUBJECTS					
		Standardised Residual		**Relative Value Added**		**Attitude to Subject**	
		feedback	control	feedback	control	feedback	control
1994	mean	**-0.14**	**0.20**	**-0.18**	**-0.02**		
	n	142	64	136	63		
	s.e.	0.08	0.11	0.06	0.09		
1995	mean	**-0.01**	**0.39**	**-0.17**	**0.06**	**3.29**	**3.34**
	n	160	76	151	76	134	68
	s.e.	0.08	0.07	0.05	0.06	0.06	0.10
1996	mean	**-0.18**	**0.02**	**-0.18**	**-0.03**		**3.52**
	n	150	144	149	144	0	124
	s.e.	0.08	0.07	0.06	0.05		0.06
1997	mean	**0.14**	**-0.04**	**0.02**	**-0.17**	**3.38**	**3.53**
	n	137	100	131	98	124	88
	s.e.	0.08	0.10	0.06	0.07	0.06	0.08

Table 30 and Table 31 show a somewhat more complex picture of the apparent effects of the feedback. Once again, differences in Attitude to Subject are not clear, so we may restrict ourselves to comments about examination performance.

In the numeric subjects, teachers in the feedback group had consistently better results in 1994-6 than the control, measured both by Standardised Residuals and by Relative Value Added. The results for 1997 continue this pattern, increasing the difference slightly. For the non-numeric teachers the difference is reversed in the years before the feedback was sent, with the feedback group having worse results every year, both in terms of Standardised Residuals and Relative Value Added. However, in 1997, after receiving the feedback, the trend was reversed and the feedback group performed better than the control on both measures. Hence it looks as though it may have been the teachers of English and French whose students gained most advantage from the feedback.

However, a number of cautions should be raised before drawing any firm conclusions from these data. Although the numbers of student results in each group are large enough to make the findings seem robust, the number of departments is small. In fact, after the removal of the departments with missing 1997 data, there were only two non-numeric departments in the feedback group and two in the control, with three in the feedback group and two in the control for the numeric departments. All these departments were drawn from just three institutions. Thus any differences between feedback and control groups would be very sensitive to any 'local' influences that may have affected a particular department in a particular year. It would be hard to rule out the possibility that some factor wholly unrelated to the effect of the feedback sent, such as changes of personnel or policy, or events such as inspection, could have influenced a whole department's performance significantly. Having said that, there were no major changes of personnel, either within the senior management of any of the institutions or within the departments themselves during the period of the experiment. None of the institutions were undergoing inspection (although one of them had only just been inspected by FEFC) and the effects of any changes in intake were to a large extent allowed for by the adjustment in the residual gain model.

When the feedback and control groups were analysed 'whole', there were probably enough departments in each group (five and four, respectively) to make it unlikely that any local effects could be wholly responsible for the difference.

Moreover, the matching of departments prior to allocation to treatment groups had guaranteed that each institution was represented about equally in each group, so any 'local' factors that would have affected the performance of the whole institution can be ruled out as explanations for the difference between the feedback and control groups. However, when they were split into Numeric and Non-numeric subjects, the number of departments was small enough that an unattributed change in the performance of a single department could have affected the outcome. Also, despite the intention to pair departments by subject before random allocation to treatments, in one large institution (Institution 4) both the numeric departments were in the feedback group and both the non-numeric were in the control. It may be seen as unfortunate (with hindsight!) that these departments were paired by size rather than subject type.

Effect sizes

Effect sizes for the differences between feedback and control groups were calculated and are shown in Table 32. The effect size was calculated in two ways: firstly, using only the outcome scores for 1997, and secondly, using the difference between the 1997 average score and the average for the previous three years. These effect sizes are referred to as 'outcome' and 'change' respectively. In both methods the difference was standardised with a pooled estimate of the standard deviation of the 1997 outcomes. As these outcomes are already 'adjusted', they have smaller variance than the raw measure of performance from which they are derived, and the effect sizes standardised against them are not comparable with effect sizes standardised against the full spread of population variation. However, the 'population' being considered here is A level candidates, who are themselves a highly restricted sample of the whole population of human beings. This issue illustrates one of the difficulties of interpreting effect sizes. Where an outcome is measured on a meaningful scale (e.g. A level grade), it is generally more useful and easier to interpret if any differences are presented in terms of that scale. The standard error for each effect size estimate is also shown (in brackets), although each standard errors are calculated on the assumption that individual students' results are independent, and are therefore likely to underestimate the true error substantially.

Table 32: Effect sizes for feedback effects on student performance and attitude

	OUTCOME MEASURE		
	Standardised Residual effect size (s.e.)	**Relative Value Added** effect size (s.e.)	**Attitude to Subject** effect size (s.e.)
All students			
'outcome'	0.30 (0.01)	0.28 (0.01)	0.06 (0.01)
'change'	0.31 (0.01)	0.21 (0.01)	0.02 (0.01)
Numeric subjects			
'outcome'	0.39 (0.02)	0.29 (0.02)	0.40 (0.02)
'change'	0.11 (0.02)	-0.20 (0.02)	-0.06 (0.02)
Non-numeric subjects			
'outcome'	0.20 (0.02)	0.27 (0.02)	-0.20 (0.02)
'change'	0.49 (0.02)	0.53 (0.02)	0.03 (0.02)

Table 32 shows that the feedback effect size was much the same whether or not 1997 results were compared with previous years' performance. In terms of 'value added' performance (i.e. A level grades, adjusted for prior attainment), the effect size for the feedback was about 0.3.

Given that there seems to be an overall tendency for the feedback group to have performed better, it is tempting to try to identify any subgroups that may have benefited particularly. Dividing the sample into numeric and non-numeric groups is one such attempt. However, there are dangers in splitting the sample, and analysing the subgroups separately, on any criteria other than those on which they were matched before random allocation. Although it may be useful to know which teachers and students performed best after receiving the feedback, it cannot necessarily be assumed that the apparent benefit was *caused* by the feedback, since the effect could be confounded with some other uncontrolled variable. Nevertheless, a number of subgroups were analysed and effect sizes calculated.

One of the factors that seemed worth investigating was whether students had been taught in a set all of whose teachers had received (or been eligible for) feedback, as opposed to those where only some of the teachers had been in the experiment. Just over half the students were in sets with all teachers participating, and the feedback

effect size for this subgroup was 0.09,[21] compared with 0.53 for those in sets where at least one teacher was not involved in the experiment. This result seems totally against what would have been expected and is quite hard to reconcile with the inference that sending feedback to their teachers caused an improvement in students' results. Surely if all their teachers had the feedback, the effect would be larger than if only some received it? Possibly in line with this finding, but equally counter-intuitive is the result that the effect size of the feedback for those students who were taught by a single teacher (just under a third of the sample) was -0.35, while for those taught by more than one teacher it was 0.48. It had been conjectured that one of the effects of sending the feedback would be to make the teachers feel more accountable for their students' performance, and in taking responsibility for it would focus on it and become more motivated to improve it. If this were the case, however, it might be expected that the feedback effect would be greater in sets taught by a single teacher than in shared groups, since in the former the responsibility would be that much clearer.

Some light was thrown on these findings when the interaction between the number of teachers who taught the group and whether or not all of them were in the experiment was examined. Of the students taught by more than one teacher, 509 were in sets where all the teachers were in the experiment, while 967 had only some of their teachers involved. The effect sizes for the feedback on these groups were 0.37 and 0.53 respectively, and these two effects are sufficiently close for it to seem that there was no significant difference between them. In other words, whether all or just some of the teachers were involved in the experiment did not really make a difference: the apparent difference in effect sizes is largely explained by the number of teachers teaching each set.[22] The feedback 'effect' was much greater on the results of students taught by several teachers than on those taught by a single teacher. The negative feedback effect for those taught by a single teacher (-0.35) is quite hard to interpret, since for this subgroup the feedback and control groups did not seem to be very well

---

[21] These effect sizes were calculated from the change in mean standardised residuals (i.e. mean for 97 – mean for 94-6). The difference between the change for the feedback group and the change for the control group was standardised by dividing by the pooled estimate of standard deviation of the 97 residuals, restricted to the particular subgroup in question.

[22] Note that all the results included for students taught by a single teacher will have had all of their teachers involved in the experiment.

matched. In the control group for 1997 there were only 39 students in 3 teaching sets, all in the same department, and their average Standardised Residual was an impressive 0.72. This compares with 133 results in 1997 for the feedback group, in 11 teaching sets in 2 departments, with an average Standardised Residual of 0.23. In fact the improvement in 1997 (compared to 1994-6) for students in the feedback group was about the same for all these subgroups, irrespective of the number of teachers they had or whether all of them were in the experiment, but the effect size varied greatly because of the changing performances of the comparable students in the control group. It therefore seems unwarranted to infer that the differences in the performance of those who had received the feedback and those who had not were attributable to the number of teachers teaching each set.

Splitting up the individual subjects also provides some interesting findings. The feedback effect sizes are 0.99 for English, 0.34 for French, 0.47 for Mathematics and -0.26 for Physics. Once again, however, some of these figures are not quite what they might seem. The huge effect in English is almost entirely explained by a poor performance by the control group in 1997 (average Standardised Residual of -0.53 for 61 students, all in the same department, compared with a pre-97 average of 0.11); the group who received the feedback improved only slightly on their pre-97 performance, but relative to the control their improvement was colossal. Equally, the negative figure in Physics is entirely attributable to the outstanding performance of one of the two departments in the feedback group in the years 1994-6, which they were unable to maintain in 1997, despite what would otherwise have been considered very good results. This department achieved an astonishing average Standardised Residual of 1.03 (with 89 students) before receiving the feedback and dropped to only(!) 0.51 (with 30 students) in 1997. The other department, with approximately the same numbers of students, averaged 0.12 before and 0.14 after.

Multilevel models

Model 1: Students within sets (2 levels)

The data for this analysis consisted of the A level grades of 443 students in 43 teaching sets in the year 1997. For this initial model it was decided to fit just two levels: students and sets. In Model 1a, no adjustment was made for any of the intake variables. In Model 1b, A level grades were adjusted for prior achievement (average GCSE score). In 1c, adjustment for prior achievement (average GCSE) and parental occupation was allowed, and finally in 1d, adjustment for sex was also included. The parameter estimates for these models are shown in Table 33. The likelihood of each model is also shown and the corresponding chi-squared probabilities of achieving such a likelihood by chance after the inclusion of each additional parameter in the model, even if there was no genuine explanatory effect of that parameter. It can be seen that, on this basis, the inclusion of average GCSE and parental occupation as explanatory variables are justified (Model 1c), but the further inclusion of sex (Model 1d) does not improve the statistical fit more than might have been expected for a purely random variable.

A fifth variation on this model was also fitted in which Model 1c was modified to allow the slopes of the A level grade/average GCSE relationship to vary between sets. However, the fixed effects coefficients were very similar to those in the corresponding model with fixed slopes, and the estimates of Level 2 variance were zero. Moreover, the likelihood value of 1947.71 for this model suggested that it fitted the data less well, despite the inclusion of an extra parameter. The parameter estimates for this model have therefore not been shown.

Table 33: Model 1: 2-level ML models

	Model 1a A level grade, unadjusted			**Model 1b** A level grade, adjusted for avg GCSE		
Fixed Effects Coefficients						
	estimate	**(s.e.)**		**estimate**	**(s.e.)**	
intercept	5.74	(0.29)		-8.51	(0.89)	
average GCSE				2.29	(0.14)	
parent occupation						
sex						
treatment	0.62	(0.38)		0.77	(0.27)	
Random Effects Variance						
	est.	**(s.e.)**	**%**	**est.**	**(s.e.)**	**%**
between sets	0.74	(0.32)	9	0.28	(0.16)	6
between students	7.18	(0.51)	91	4.57	(0.32)	94
Goodness of Fit						
-2loglikelihood	2160.09			1950.46		
p(improved)				0.0000		

	Model 1c A level grade, adjusted for avg GCSE and parental occupation			**Model 1d** A level grade, adjusted for GCSE, parental occupation and sex		
Fixed Effects Coefficients						
	estimate	**(s.e.)**		**estimate**	**(s.e.)**	
intercept	-9.48	(0.95)		-9.45	(0.95)	
average GCSE	2.23	(0.14)		2.24	(0.14)	
parent occupation	0.30	(0.10)		0.30	(0.10)	
sex				-0.07	(0.22)	
treatment	0.75	(0.26)		0.75	(0.27)	
Random Effects Variance						
	est.	**(s.e.)**	**%**	**est.**	**(s.e.)**	**%**
between sets	0.25	(0.15)	5	0.26	(0.15)	5
between students	4.50	(0.32)	95	4.49	(0.32)	95
Goodness of Fit						
-2loglikelihood	1942.51			1942.42		
p(improved)	0.0048			0.76		

Note: 'treatment' is a dummy variable, taking the value 1 for results of students in the feedback group, 0 for the control group. Its coefficient is therefore an estimate of the average difference between the two groups, after adjustment for all the other variables included in that model. 'p(improved)' is an estimate of the probability that the improvement in fit over the previous model could have arisen by chance (=$\chi^2_{(\nu)}$ (change in -2loglikelihood), where ν = no of additional parameters)

Statistically, the best fit for Model 1 is 1c, in which A level grades have been adjusted for average GCSE score and parental occupation. Although the coefficient of 'parental occupation' is not large, it is large enough to be statistically significantly different from zero. Its inclusion reduces the total residual variance by just 2% (from 4.85 in Model 1b to 4.75 in 1c). The coefficient of 'sex' in 1d is small, both in absolute terms and relative to its standard error, and its inclusion makes almost no difference to any of the parameter estimates anyway.

Clearly, average GCSE score is the most important explanatory variable and once adjustment for this has been made (i.e. in Models 1b, 1c and 1d) the estimates for the effect of the feedback remain fairly stable. The value of 0.75 for the coefficient of 'treatment' (Model 1c) suggests that the adjusted A level grades of students in the feedback group were 0.38 of a grade better than those in the control, i.e., a feedback improvement effect of well over one third of a grade. This coefficient is roughly three times its standard error and therefore highly statistically significant. Using 3.23 as the pooled estimate of the standard deviation of 1997 A level grades (derived from the data in Table 28), this corresponds to an effect size of 0.23.

One of the surprising features of Model 1 is the very small amount of between-sets variance. Just 5% of the variance was between sets (i.e. a within-set correlation of only 0.05). This figure is hard to explain, but suggests that residuals calculated from multilevel models and those that ignore the multilevel structure (i.e. OLS residuals, as used by ALIS) are likely to be indistinguishable.

Model 2: Students within sets, within departments (3 levels)

The second set of multilevel models fitted allowed three levels of the hierarchy: students within sets, within departments. The 443 students in 43 sets were therefore now recognised as coming from 9 departments. Four versions of this model were once again fitted, with the same progression of explanatory variables as in Model 1.

Table 34: Model 2: 3-level ML models

	Model 2a A level grade, unadjusted			**Model 2b** A level grade, adjusted for GCSE		
Fixed Effects Coefficients						
	estimate	**(s.e.)**		**estimate**	**(s.e.)**	
intercept	5.89	(0.45)		-8.64	(0.93)	
average GCSE				2.32	(0.14)	
parent occupation						
sex						
treatment	0.31	(0.62)		0.64	(0.38)	
Random Effects Variance						
	est.	**(s.e.)**	**%**	**est.**	**(s.e.)**	**%**
between depts	0.59	(0.40)	7	0.17	(0.15)	3
between sets	0.27	(0.24)	3	0.13	(0.14)	3
between students	7.16	(0.50)	89	4.57	(0.32)	94
Goodness of Fit						
-2loglikelihood	2154.03			1948.17		
p(improved)				0.0000		

	Model 2c A level grade, adjusted for GCSE and parental occupation			**Model 2d** A level grade, adjusted for GCSE, parental occupation and sex		
Fixed Effects Coefficients						
	estimate	**(s.e.)**		**estimate**	**(s.e.)**	
intercept	-9.62	(0.45)		-9.53	(1.00)	
average GCSE	2.26	(0.14)		2.29	(0.15)	
parent occupation	0.29	(0.10)		0.30	(0.10)	
sex				-0.21	(0.23)	
treatment	0.63	(0.37)		0.61	(0.40)	
Random Effects Variance						
	est.	**(s.e.)**	**%**	**est.**	**(s.e.)**	**%**
between depts	0.16	(0.14)	3	0.21	(0.16)	4
between sets	0.11	(0.13)	2	0.11	(0.13)	2
between students	4.50	(0.32)	94	4.48	(0.31)	93
Goodness of Fit						
-2loglikelihood	1940.22			1939.48		
p(improved)	0.0048			0.39		

The four versions of Model 2 show much the same pattern as found in Model 1, with version 'c' being the best fit statistically. The fixed effects coefficients have not changed much from Model 1 and the differences are certainly well within their statistical margins for error. The estimate for the treatment effect has dropped slightly, however, to 0.63 (i.e. just under a third of a grade, with a corresponding

effect size of 0.19) and its standard error has increased appreciably. This increase in standard error is what would have been expected, since we no longer have to repeat the caveat that standard errors are underestimates because they assume the independence of teachers within the same department. In the multilevel model used here, the similarity of performance of students within the same set and of sets within the same department is explicitly modelled, and the standard error is no longer an underestimate. Unfortunately, however, the estimate of the treatment effect is now not large enough to provide conventional justification for the rejection of the 'null' hypothesis of no effect (in fact, on a non-directional test, $p = 0.09$).

Once again, the overwhelming majority of the variance is between students, with only very small percentages between departments and between sets. However, the standard errors for these variance estimates are large enough to make the true proportions rather uncertain.

Model 3: Different subject coefficients

One of the founding principles of the ALIS project was that different A level subjects should be modelled separately, in other words that the relationship between average GCSE and A level grade be allowed to vary across subjects. The justification for this lies in the fact that different A levels have always catered for quite different populations of candidates and have used widely different assessment procedures. The same grade in two different subjects therefore represents two quite different achievements, and the difference should be recognised in any calculation of value added. The corollary of this, that some subjects are more 'difficult' (i.e. more severely graded) than others, has been robustly defended (Fitz-Gibbon and Vincent, 1994, 1997) despite some criticism (Goldstein and Cresswell, 1996).

In Model 3, the fixed regression coefficients were allowed to vary across subjects by including four dummy variables (English, French, Mathematics, Physics) each of which took the value 1 if the result was in that subject, and 0 otherwise. A further four variables were created, one for each subject, which took the value of the average GCSE score if the result was in that subject, and 0 otherwise. Thus the regression coefficients associated with the two variables for each subject would be restricted to the cases in that subject. The model was once again fitted using also the

dummy variable 'treatment' (1 if the case was in the feedback group, 0 if in the control group) and parental occupation. The parameter and variance estimates are shown in Table 35, as are the regression coefficients in each subject for the 1997 ALIS cohorts. In comparing these ALIS coefficients with the fixed effects estimates from the multilevel model it should be noted that the ALIS models do not incorporate parental occupation and also that some of the subjects are themselves subdivided (e.g. English into Language and Literature, Mathematics into different syllabuses and Physics into modular and non-modular courses) with different regression equations for each.

Table 35: Model 3: Different subject coefficients, 2 level and 3 level ML models

		Model 3a 2 levels: students, sets			**Model 3b** 3 levels: students, sets, depts			**ALIS coefficients** (OLS)
Fixed Effects Coefficients								
		estimate	**(s.e.)**		**estimate**	**(s.e.)**		
English	Intercept	-9.07	(1.30)		-9.17	(1.30)		-8.44
	average GCSE	2.18	(0.19)		2.19	(0.19)		2.42
French	Intercept	-6.02	(2.49)		-6.02	(2.51)		-14.17
	average GCSE	1.83	(0.37)		1.76	(0.37)		3.07
Maths	Intercept	-7.56	(1.84)		-7.48	(1.84)		-11.35
	average GCSE	1.90	(0.30)		1.88	(0.30)		2.71
Physics	Intercept	-13.46	(1.98)		-13.68	(1.99)		-12.62
	average GCSE	2.80	(0.31)		2.83	(0.31)		2.88
	parental occupation	0.25	(0.10)		0.26	(0.10)		
	treatment	1.06	(0.27)		1.12	(0.28)		
Random Effects Variance								
		est.	**(s.e.)**	**%**	**est.**	**(s.e.)**	**%**	
	depts				0	0	0	
English	sets	0.14	0.16	4	0.14	0.16	4	
	students	3.46	0.38	96	3.47	0.26	96	
	depts				0.89	1.09	2	
French	sets	0.61	0.61	18	0.03	0.26	0	
	students	2.76	0.60	82	2.76	0.60	98	
	depts				0	0	0	
Maths	sets	0	0	0	0	0	0	
	students	6.22	0.78	100	6.23	0	100	
	depts				0.05	0.27		
Physics	sets	0.37	0.39	8	0.33	0.44		
	students	4.42	0.68	92	4.41	0.67		
Likelihood		1910.57			1912.58			

It can be seen that the fixed effects estimates (intercepts and slopes) are very close in the two level model (3a) to those in the three level model (3b), but they vary appreciably across the four subjects, suggesting that the relationships between A level grade and average GCSE score were not the same in each subject. In two of the subjects (English and Physics) both the coefficients (in both models) are close enough to the ALIS coefficients to have plausibly been sampled from the same population, but in the other two (French and Mathematics) they are not. These large differences

suggest that in French and Mathematics at least, the students in the experiment were not typical of those in the whole ALIS cohort with respect to the relationship between their A level grades and prior achievement. They also suggest that, although it may be desirable in principle to model different subject 'difficulties' with different regression equations, the differences in the subject coefficients for this sample do not reflect the generally found pattern of differential difficulty in the wider cohort. In other words, differences in subject difficulty may be real, but they are not responsible for the variation in subject regression coefficients found in model 3. Although, in principle, different subjects should be modelled separately, because of the small size and lack of representativeness of this sample, results from this model may well not generalise to a larger population.

The percentages of variance associated with each level of the model also vary with the subject, but the standard errors of these variances are again sufficiently large to make the proportions somewhat uncertain. In terms of the goodness of fit for each model, when the likelihood for model 3a is compared with model 1c on a chi-squared test with 9 degrees of freedom (instead of the 4 explanatory variables and 1 variable whose variance is estimated at level 2 in model 1c, we have 10 explanatory variables and 4 variances in model 3a), it is found to improve the fit well beyond what would have been expected by chance (in fact, $p= 6\times10^{-5}$). Although the likelihood of model 3b has risen from model 3a despite incorporating additional parameters (the extra level requires the estimation of variances for 4 further variables), it is still a sufficiently better fit than model 2c to reject the explanation that the improvement is attributable to chance ($p = 0.01$). Hence, in purely statistical terms, both these models may be considered a better fit than models 1 and 2.

However, in using a separate regression equation for each subject, we have effectively paired off the feedback and control groups within each subject, and may thus once again be in danger of exploiting differences between subgroups of the sample that were not properly matched before random allocation (see discussion on p242). If residuals in different subjects are calculated from different equations, they are effectively weighted unequally. If the different subject coefficients vary only in their intercepts, then it will make no difference to the calculation of the average treatment effect. However, if the slopes also vary (as they do in model 3), then the

calculation of the average treatment effect could be sensitive to any initial differences between the feedback and control groups within a given subject.

In both these models the coefficient of 'treatment' is greater than one, indicating that the students in the feedback group achieved over half a grade better than those in the control, after the 'effects' of other factors were controlled. These coefficients are around four times their standard errors, so are well above what would be required to reject a null hypothesis at any reasonable level (in fact $p < 5\times10^{-5}$ for both). The estimates of the treatment effect in models 3a and 3b correspond to effect sizes of 0.33 and 0.35 respectively. These are quite substantial effects, but for the reasons given above they may not be very robust.

Model 4: Adjustment for previous departmental performance

The data for the years 1994 to 1996 were used to estimate a residual score for each department in the experiment. A three level model was fitted to these data with students within years within departments, and A level grade was adjusted for average GCSE score and parental occupation. The departmental residuals varied from roughly -2 to 2 (i.e. all departments averaged within one grade of expectation), with an average of 154 students each. However, the residuals for departments with over 100 students (6 of them) all had absolute value less than 1. The residual for each of the 9 departments in the experiment was then entered as an explanatory variable in model 4, in addition to the variables already used in model 2c. One of the possible advantages of using past performance as an explanatory variable was thought to be that it would reduce the need to model each subject separately. If one subject were consistently 'harder' than another, departments in that subject might be expected to have lower residuals for 1994-6, and their apparently less good performance in 1997 would be adjusted to take account of this effect, provided the relative difficulty of subjects was stable over time. In a sense, therefore, model 4 is a compromise between the need to model different subjects separately and the potential instability of doing this with a small sample. The parameter estimates for model 4 are shown in Table 36.

Table 36: Model 4 Adjustment for previous departmental performance

	Model 4 A level grade, adjusted for GCSE, parental occupation and dept's previous residual		
Fixed Effects Coefficients			
	estimate	**(s.e.)**	
intercept	-9.87	(0.99)	
average GCSE	2.28	(0.14)	
parent occupation	0.30	(0.10)	
dept's prev. resid	0.35	(0.23)	
treatment	0.68	(0.33)	
Random Effects Variance			
	est.	**(s.e.)**	**%**
between depts	0.10	(0.11)	2
between sets	0.10	(0.13)	2
between students	4.50	(0.32)	96
Goodness of Fit			
-2loglikelihood	1938.12		

Interestingly, the inclusion of the department's residual from the previous three years did not improve the fit of the model beyond what might have been expected by chance ($p = 0.15$), and its coefficient in the fixed effects part of the model was not large enough (relative to its standard error) to be considered statistically significantly different from zero. This indicates that knowledge of the department in which a student was taught did not significantly improve the prediction one could make for that individual student's A level performance. However, the regression coefficient of 0.35 suggests that, on average, students in a department with a good previous performance did appear to benefit by approximately one third of the previous residual. In a department whose previous results were, say, half a grade better than the norm, one could therefore predict that next year's average would be about one sixth of a grade better than the background population. Compared to the variation in individual performance within the department, however, this advantage would be too small to enable appreciably better individual predictions to be made.

The coefficient of the treatment dummy (i.e. the average difference between the performance of those in the feedback and control groups) has increased slightly and its standard error has reduced compared with model 2c. It is now above the magic

threshold of statistical significance. However, the size of the feedback effect remains at about a third of a grade, corresponding to an effect size of 0.21.[23]

Analysis by teachers

Analysing the results of the individual students in the feedback and control groups (as above) provided an indication of the effects of the feedback on the students. However, the feedback was actually given to the teachers. It was therefore also hoped to be able to get an indication of the effects on the teachers, and to investigate any associations between feedback effects and characteristics of the teachers.

The analysis using multilevel models (above) shows that the proportion of variance within sets and within departments was generally quite small, and frequently not statistically significantly different from zero. This finding made it hard to justify continuing the analysis using multilevel models in preference to ALIS's OLS residuals, given the desire to model different subjects separately and the anxieties about the robustness of individual subject regression equations derived from this sample. Moreover, the simplicity and transparency of the OLS residuals seemed to favour them.

A 'teacher average' for each of the three outcome measures (Standardised Residual, Relative Value Added and Attitude to Subject) was calculated by taking all the students taught by that teacher and weighting the outcomes by the proportion of the teaching for which that teacher was responsible. Thus if a teacher taught one set alone and shared another equally with a colleague, the results of the students in the former set would have twice the weight in the calculation of their average. The assumption underlying this calculation was that teachers in shared groups had 'linear' effects on their students, that is to say that the total teacher effect on a student in a

[23] This effect size is calculated as the difference between the averages for the two groups, divided by the standard deviation of the raw A level scores. If, instead, it were calculated by dividing by the standard deviation of the residuals (2.17 in this case), the effect size would rise to 0.31. This difference illustrates a difficulty in interpreting effect sizes. If the effect of the feedback is considered in terms of *performance* (i.e. A level grades) then the former value is appropriate; if, however, it is the effect in terms of *progress* (i.e. performance relative to starting point, or the residual) that is required, then the latter would be more appropriate. If, on the other hand, one wishes simply to make

shared group was the sum of the individual teacher effects. Clearly this assumption is inadequate, since teachers are likely to interact with each other and some combinations may well be more (or less) than the sum of their parts. However, it provided a convenient simplification and may be viewed as an acceptable starting point for modelling the effects of teachers on their students' performance.

Stability of teacher averages

It is a commonly made assumption in school effectiveness research that 'value added' provides a measure of 'effectiveness' (see Chapter 2). Moreover, it has increasingly been realised that it is at the level of the classroom that effects should be sought (Creemers, 1994; Hill and Rowe, 1996). It therefore came as something of a surprise – and no small disappointment – to discover that no previously recorded reports of the stability of teacher 'effects' could be found. Evidence about the stability of value added scores would provide a crucial test of the hypothesis that they measure 'effectiveness'. If they are found to vary wildly from year to year for the same teacher, they could not really be seen as a reliable measure of effectiveness. If, on the other hand, they are reasonably stable – or better still if any 'instability' can be accounted for by other factors, or as part of a trend – then one's faith in their interpretation as 'effectiveness' would be strengthened.

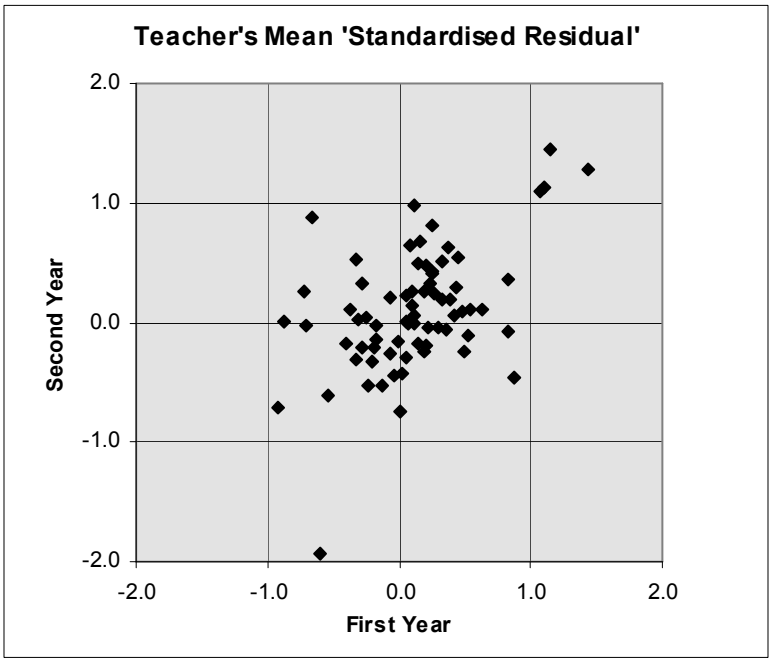
With the data from this study, an estimate of the stability of teacher 'effectiveness' was calculated as follows. For each teacher, an average for all the students they had taught in each year was calculated for each of the variables Standardised Residual, Relative Value Added and Attitude to Subject. Where a teacher had an average in two successive years that could not have been influenced by the experimental feedback (i.e. for pairs of averages across 1994-5, 1995-6 and for the control group in 1996-7) the pairs were used to calculate an estimate of the correlation between a teacher's average in one year and the next. These pairs are plotted in the three scatter graphs in Figure 9. It should be remembered that the pairs are not independent, since a given value may appear twice: once as the first year of a pair and

---

one's results seem as impressive as possible, then it is not difficult to provide convincing justification for whichever value is larger.

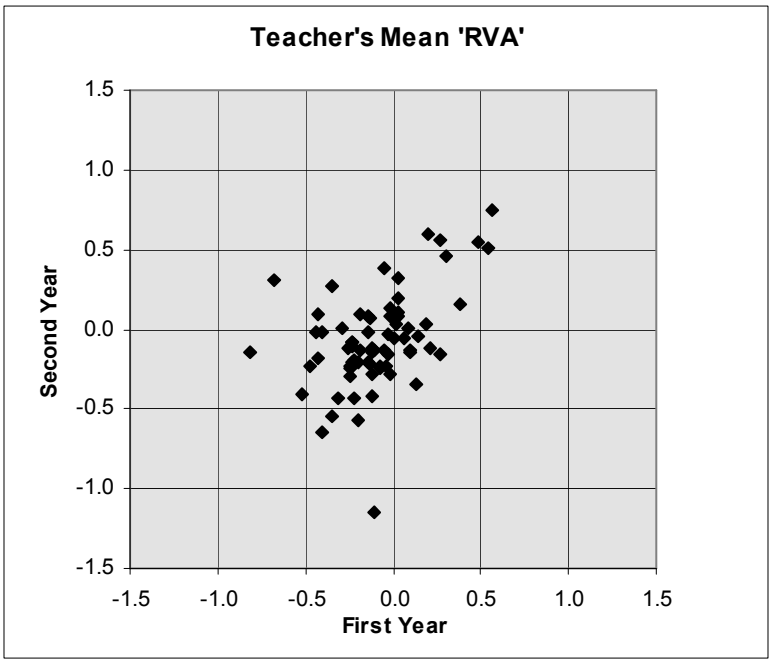
again as the second year of another pair.[24] Hence the confidence intervals quoted (which are based on the assumption of independence) are likely to underestimate the range in which the 'true' value may be expected to lie. However, assuming that the sample is representative, the value of the correlation coefficient calculated in this way is an unbiased estimate of the population correlation.

Figure 9: Year on year correlation for teacher averages

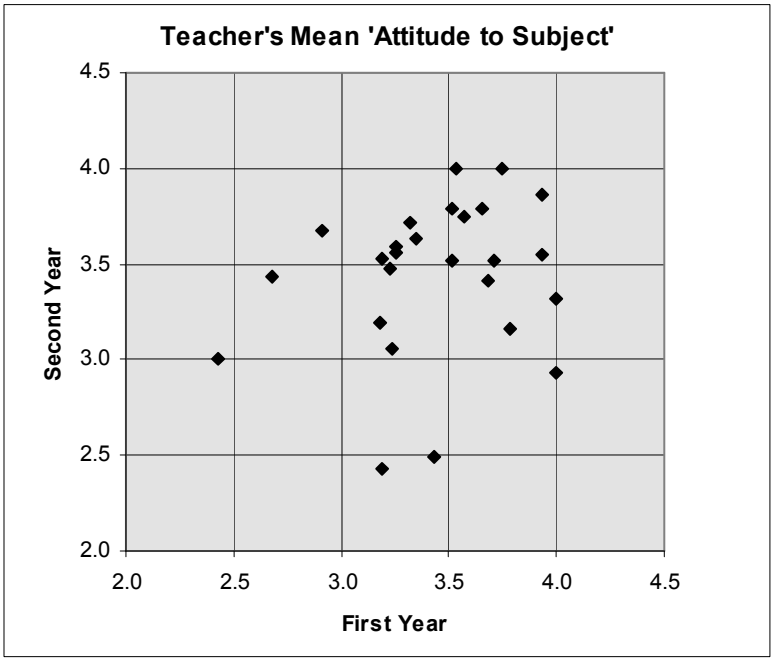


Correlation = 0.50, n = 71, 95% CI = [0.30, 0.66]

[24] They may also fail to be independent if part of the stability in a teacher's average from year to year is rightly attributable to the effect of being in the same department, i.e. if the results of teachers in the same department are more similar than those in different departments. This is a further example of the 'clustering' issue mentioned previously. The use of multilevel models could provide a theoretical solution to the problem, but the small numbers involved here would mean that the standard errors of any parameters estimated would be too large to make it of much practical benefit.



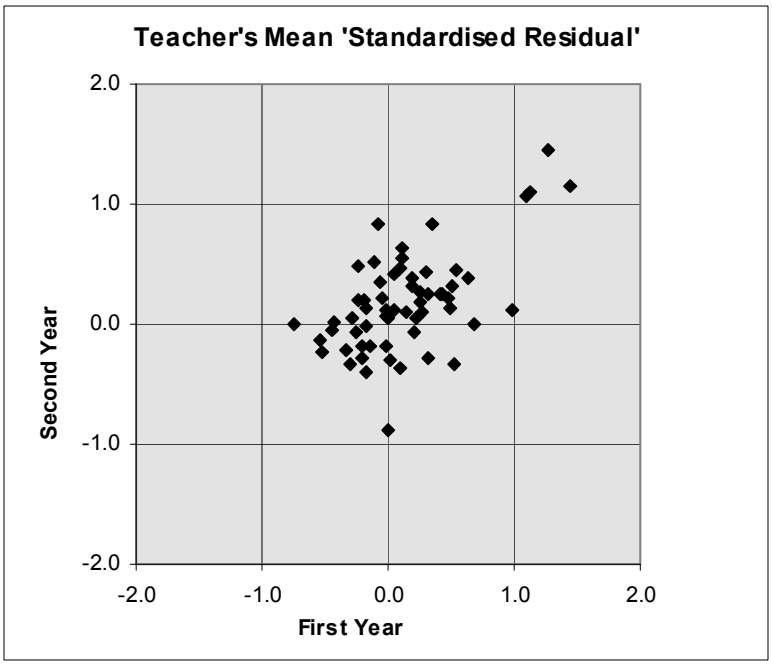
Correlation = 0.48, n = 71, 95% CI = [0.27, 0.64]



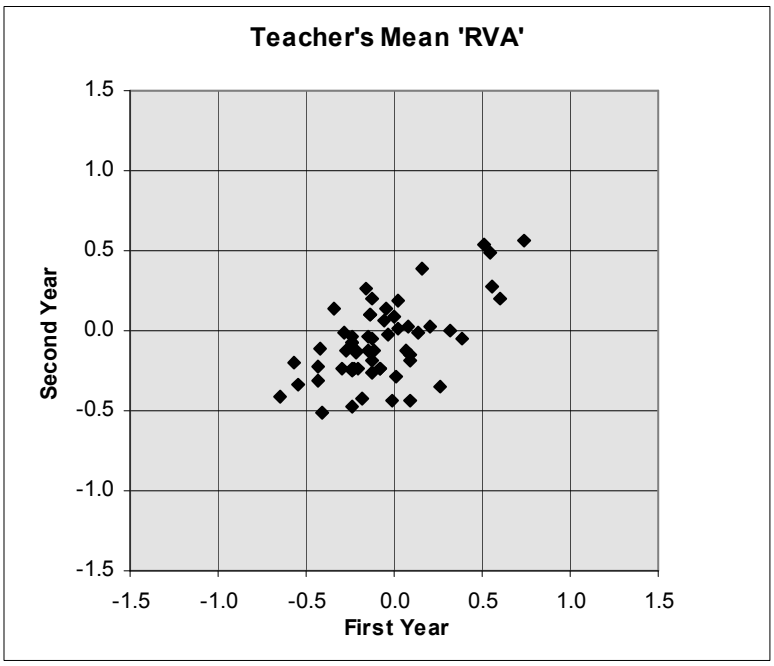
Correlation = 0.20, n = 29, 95% CI = [-0.18, 0.53]

The same calculation of the correlation coefficient and the plotting of the graphs was then repeated, but for a sub-sample restricted to those teacher averages which comprised at least five (or the equivalent: for example, ten students shared by two teachers) students. These graphs are shown in Figure 10.

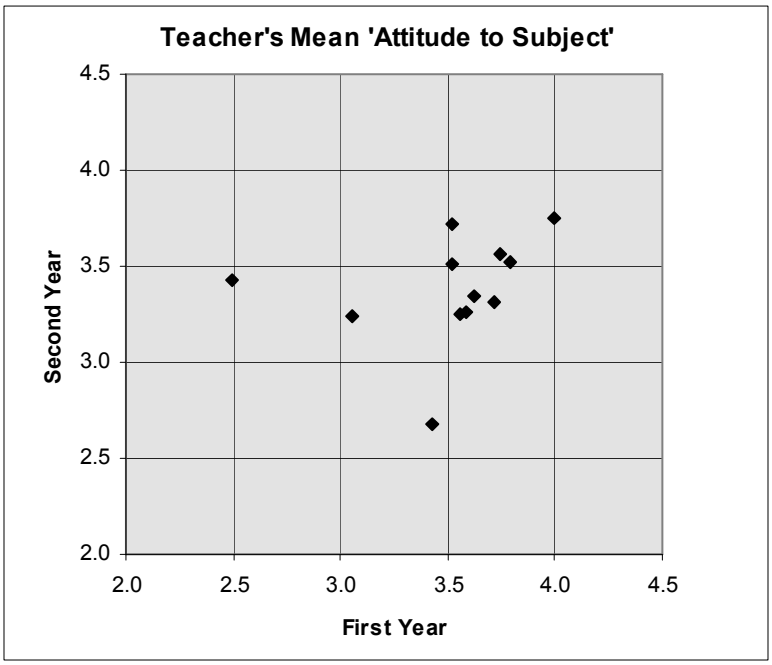
Figure 10: Year on year correlation for teacher averages, restricted to averages of ≥ 5 students



Correlation = 0.61, n = 59, 95% CI = [0.42, 0.75]



Correlation = 0.61, n = 57, 95% CI = [0.42, 0.75]



Correlation = 0.25, n = 12, 95% CI = [-0.38, 0.72]

For the 'performance' variables (Standardised Residual and Relative Value Added) the restriction to averages of at least five students appears to have increased the correlation (although not enough for the difference to be statistically significant) and to have removed a few 'stray' pairs from the graph. Interestingly, although the correlations for Standardised Residual and Relative Value Added are equal (0.61), the graphs of the latter suggests more strongly a general relationship between the averages in successive years, while the correlation coefficient in the former appears to depend more on the presence of four outliers in the top right of the graph. In fact, these four pairs are the results for just two extremely successful teachers with results in 1994-5 and 1995-6. When the four points are removed, the impression given by the remainder is of a much weaker relationship and, indeed, the correlation drops without them to just 0.30.

The strength of the correlation between successive years' averages of Relative Value Added (RVA) provides some justification for its inclusion. It was hoped that this measure would be more sensitive to the effect of an individual teacher since it excluded that part of a student's value added that was common to all their subjects, and this hope is to some extent encouraged by the evidence from the graphs. It must be acknowledged, however, that RVA as it was calculated here (the difference between value added in the subject in question and the average for all subjects taken)

was a crude attempt to adjust for the correlation between a given student's value added in different subjects. Further investigation would be required to establish the amount and type of adjustment necessary to optimise the stability of the measure.

The most that can be said about the stability of teacher averages of Attitude to Subject is that the numbers in this sample are too small to be able to make much of an estimate. Certainly, the correlation between attitudes of students taught by the same teacher in successive years does not seem to be very high, and may indeed be close to zero. This suggests that student attitudes may not have been significantly influenced by the teacher, although it would be necessary to look at a much bigger sample to conclude this with any confidence.

This analysis of stability is a somewhat inadequate first attempt. A more sophisticated attempt might use multilevel models, which theoretically offer the opportunity to model the sharing of groups better using cross-classified models, and to estimate the stability of a teacher residual from the intra-class correlation, rho. However, as mentioned before, within the constraints of this study the cross-classified model could not be made to work and the size of the sample raised some doubts about the robustness of the ML estimates.

Despite these limitations, though, the results are quite interesting. Correlations of the order of 0.6 are high enough to suggest that teachers whose students have performed well in one year can often be expected to produce better than average results in the next. This seems to lend weight to the belief that in general teachers do have some causal effect on student performance – a belief that is so fundamental to educational practice that to challenge it would be unthinkable. However, the size of the teacher effect is perhaps disappointingly small. Certainly, a correlation of 0.6 is not high enough to justify using value added as a measure of an individual teacher's performance.[25] The reliability required for such a judgement would surely be of the order of at least 0.9. Of course, it may be that some teachers are more consistent in their 'effects' than others and that for some sub-groups the correlation would be higher. Also, by restricting it to teachers with even larger numbers of students, a higher correlation could certainly be achieved.

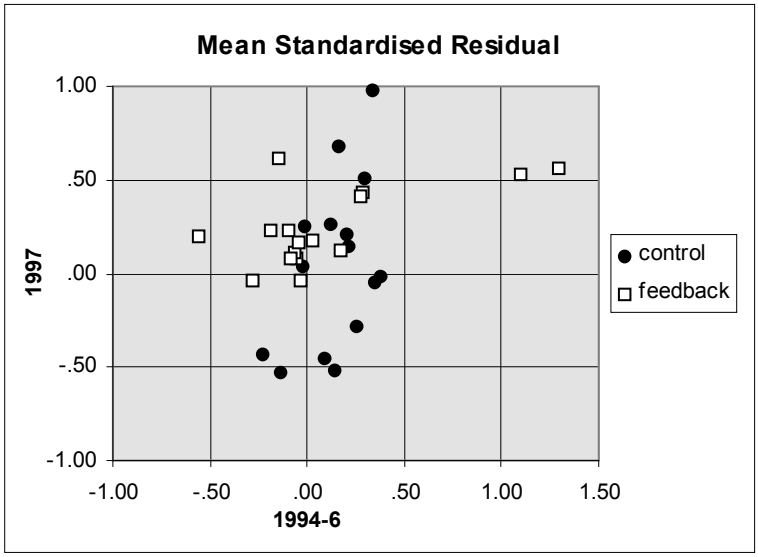
---

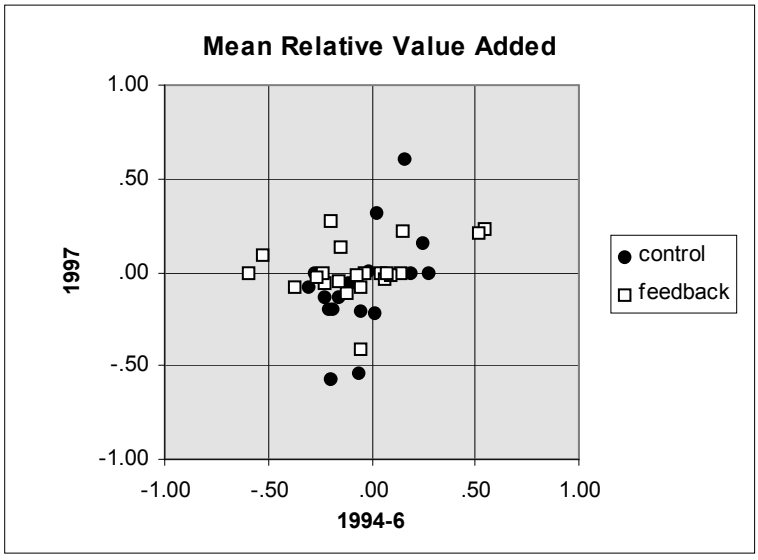
[25] With a correlation coefficient, r, the proportion of variance in one variable accounted for by the other is r^2. Hence in this example only about a third of the variation in teachers' value added scores in a given year is explicable in terms of their previous scores.

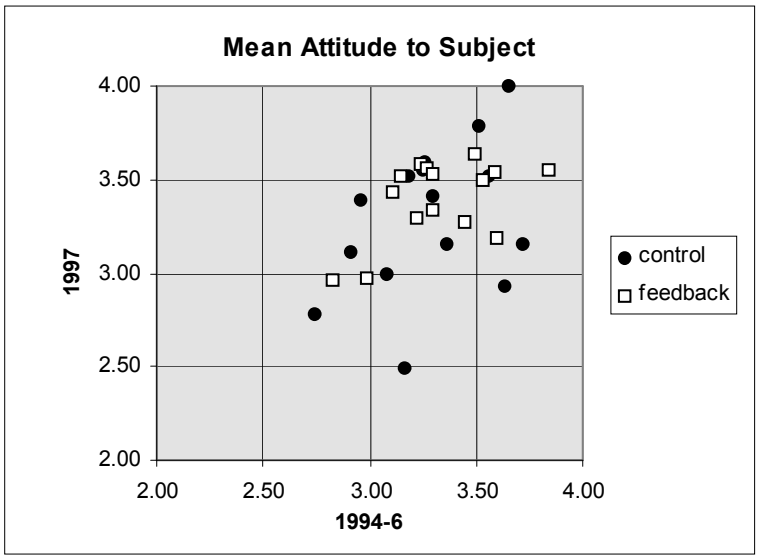
Feedback Effects

For each teacher, the average for all the students they had taught prior to the experiment (1994-6) and an average for 1997 was calculated for each of the variables. Scatter graphs for the two measures on each variable are shown in Figure 11, with the two treatment groups separated.

Figure 11: Scatter graphs for changes in teacher averages







The overall correlation coefficients for the teacher averages before and after the intervention are 0.36 for Standardised Residual, 0.39 for Relative Value Added and 0.44 for Attitude to Subject. The correlations when restricted to the control group are 0.48, 0.59 and 0.41 respectively. The increases in the correlations for the examination performance variables (i.e. Standardised Residual and Relative Value Added) when restricted to the control group help to confirm the impression from the graphs that the changes for the feedback and control groups are in fact different. Once again, however, the two treatment groups appear indistinguishable in terms of changes in Attitude to Subject, and this third graph will not be considered further.

The pattern of change for the control group is similar on both the 'performance' graphs. As might be expected, 1997 averages are more widely spread than the averages for 1994-6. The latter are based typically on considerably more students and would be expected to be less subject to the 'random' variations found when taking an average of a smaller number of values, only part of whose variation is attributable to the teacher. However, the reverse seems to be the case for the feedback group. It is as if the effect of the feedback were to reduce the spread of performance, making all the teachers closer to the overall average.

The graph of Mean Standardised Residual shows the clearest difference between the teachers in the feedback and control groups. If a diagonal is drawn along the line of equal performance before and after the feedback (i.e. from (-1,-1), through (0,0) and (1,1)), it can be seen that seven of the control group are plotted well below the line (i.e. their average in 1997 was worse than in 1994-6) and one more is just below.

However, apart from two outliers whose pre-97 averages were so high it would have been extraordinary if they had not declined, all of the feedback group are plotted close to or above the line.

Table 37: Teacher averages before and after receiving feedback

YEAR		TEACHER AVERAGE					
		Standardised Residual		**Relative Value Added**		**Attitude to Subject**	
		feedback	control	feedback	control	feedback	control
1994-6	mean	**0.10**	**0.14**	**-0.05**	**-0.08**	**3.32**	**3.28**
	n	16	15	16	15	15	15
	std. dev.	0.48	0.18	0.28	0.15	0.26	0.29
1997	mean	**0.24**	**0.05**	**0.02**	**-0.09**	**3.37**	**3.29**
	n	16	15	16	15	16	15
	std. dev.	0.21	0.45	0.17	0.29	0.22	0.40

Table 37 shows the means and standard deviations of the three teacher averages. Only in terms of Standardised Residual is there an apparently significant feedback effect.

Effect sizes for feedback effects on teachers

Effect sizes were once again calculated for each teacher's Mean Standardised Residual, using the difference in 1997 outcomes, the difference in change scores (i.e 1997 teacher average – 1994-6 average) and the difference in residual gain (from the regression of 1997 average on 1994-6 average). The pooled estimate of the standard deviation of 1997 averages (0.35 for Standard Deviation, 0.24 for RVA) was used to standardise the effect sizes. These values, together with the actual differences between the two groups, and the approximate equivalent in A level grades, are shown in Table 38.

Table 38: Differences between teacher averages for feedback and control groups

STANDARDISED RESIDUAL	Effect Size (& std err)	Difference (fbk – ctrl)	Grade Equivalent
1997 Average	0.53 (0.37)	0.19	0.25
Change Score	0.65 (0.37)	0.23	0.30
Residual Gain	0.57 (0.37)	0.20	0.27

RELATIVE VALUE ADDED	Effect Size (& std err)
1997 Average	0.43 (0.36)
Change Score	0.32 (0.36)
Residual Gain	0.38 (0.36)

It can be seen that the feedback effect on teachers' average Standardised Residual is estimated at an improvement of around a quarter of an A level grade. In terms of effect sizes (where the population spread against which the differences are standardised is the population of teacher averages) the figures are all close to 0.6. The effect sizes for Relative Value Added are somewhat smaller, being around 0.4. Both these effects would be considered quite substantial, if replicated. However, because the sample is small, the standard errors of these estimates are large and none of the effect sizes is large enough to reject a traditional null hypothesis of 'no effect'.

The question of whether the effect sizes varied for different sub-groups of teachers was considered, and some interesting variations were found. However, given the small numbers of teachers involved, and the fact that they were not matched on any of these criteria before random allocation, one must be very cautious about assuming that any of these variations would be found in other populations. Nevertheless, the following patterns may legitimately be reported as having been found in this particular sample, and the question of how far they can be generalised (if at all) must be left to further enquiry.

Firstly, there appeared to be a tendency for the effect of the feedback to be greater when restricted to teachers whose prior performance had been worse. This, of course, is not simply a 'regression effect' (the tendency of unreliably measured variables to regress towards the mean when they are re-measured), since this would presumably apply equally to both control and feedback groups – provided they started

with similar values. The effect sizes for each of the six outcome measures when the sample was restricted to those teachers whose pre-experiment averages (Standardised Residual or RVA) were negative are shown in Table 39.

Table 39: Effect size estimates for feedback effect on teachers with previously below average performance.

OUTCOME MEASURE	**Teachers with 1994-6 Std Res < 0** (fbk n=10, ctrl n=4) Effect Size (std err)	**Teachers with 1994-6 RVA < 0** (fbk n=11, ctrl n=11) Effect Size (std err)
Standardised Residual		
1997 Average	0.90 (0.62)	0.87 (0.45)
Change Score	1.04 (0.62)	1.37 (0.47)
Residual Gain	0.95 (0.62)	1.05 (0.45)
RVA		
1997 Average	0.98 (0.62)	0.67 (0.44)
Change Score	1.06 (0.62)	0.88 (0.45)
Residual Gain	1.02 (0.62)	0.76 (0.44)

The second apparent difference in feedback effect was between those who answered 'yes' to the question asking whether they thought ALIS should routinely send class-by-class feedback to the teachers involved and those who said 'no'. This question was answered by both feedback and control groups after the former had received the feedback, but before the 1997 examinations. Again, it must be remembered that the teachers in the two groups were not matched with respect to their views about who should receive the feedback before allocation to treatments, and the fact of receiving the feedback (or not) may well have influenced their answers (although no significant differences were found between them, see p172). However, it is interesting that the effect of the feedback seemed to be appreciably greater for those who stated that teachers should routinely receive this feedback than for those who said they should not (see Table 40). The differences, though, are not large enough to be statistically significant with the small numbers involved. Moreover, it is impossible to rule out the explanation that the two groups in each category were not equivalent in

some significant way before the feedback was provided, and the finding must be treated as at best suggestive.

Table 40: Effect size estimates for teachers separated by opinions as to whether ALIS should send class-by-class feedback

OUTCOME MEASURE	SHOULD ALIS SEND CLASS-BY-CLASS FEEDBACK TO THE TEACHER(S) INVOLVED?	
	Teachers who said 'yes' (fbk n=4, ctrl n=6)	**Teachers who said 'no'** (fbk n=5, ctrl n=4)
	Effect Size (std err)	Effect Size (std err)
Standardised Residual		
1997 Average	0.86 (0.67)	0.40 (0.68)
Change Score	1.45 (0.72)	-0.16 (0.67)
Residual Gain	1.07 (0.69)	0.20 (0.67)
RVA		
1997 Average	0.74 (0.67)	0.51 (0.68)
Change Score	1.02 (0.68)	-0.30 (0.67)
Residual Gain	0.86 (0.67)	0.17 (0.67)

Chapter 7

Project 2: Data Collection

7.1 CHOICE OF SAMPLE

A list of all the institutions in the ALIS project in both 1996 and 1997 was compiled, and from it the names of all those that had previously been contacted in connection with Project 1 were deleted. This left 407 institutions[26] from which a random sample of 192 was selected using SPSS. The same four subjects (English, French, Mathematics and Physics) as in Project 1 were used in order to ensure comparability, and in each institution the four departments were randomly allocated to the following four groups (the allocation was again made using the 'select random sample' command in SPSS and was constrained so that each of the 24 possible combination of subjects and treatment groups was achieved exactly 8 times):

1. **Departmental Information**. The department was sent a printout of their 1996 exam entries and their 'residuals' (i.e. value added exam performance), with any 'over' and 'under' achievers identified. Averages were calculated separately for high/low ability and male/female subgroups. They also received a list of 1997 entries with 'predicted' grades, i.e. the point score 'prediction' and the minimum grade they would need to achieve in order to gain a positive residual. 'Predicted' grades were based on the previous years' regression equation for that

[26]Some institutions which had been members of the SHA project (now incorporated into ALIS) were not included since their details were contained in a separate database.

subject. This information was accompanied by a set of notes to make it easier to interpret and some suggestions about how they might use it.

2. **Analysis by Teacher**. Each department was sent a list of their entries for 1996 and 1997 with the offer that if they sent it back with the teaching groups (and teacher) identified, I would provide an individual analysis for each teacher. They were given the option to have each teacher's feedback in a separate sealed envelope and asked to obtain the consent of all involved before replying.
3. **TAMIS**. The department was sent a TAMIS (Target Setting and Monitoring Information System) disk and instructions on how to use it. The disk contained a spreadsheet with each subject's regression equations built in, so that predicted grades and residuals could be calculated automatically. A randomly selected half of the departments also received the offer of a telephone helpline.
4. **Control**. These departments received nothing.

7.2 DISPATCH OF FEEDBACK

Heads of departments selected for 'Departmental Information' received the appropriate printout(s) and notes together with a covering letter (see Appendix 7A, 7B and 7C, pp311-315). Departments in the 'Analysis by Teacher' group received a list of exam entries for their department in 1996 and 1997 and were invited to return the lists, having indicated the teaching set for each student (see Appendix 7D, p317). Those receiving 'TAMIS' had a copy of the generic TAMIS disk, the instructions and notes and a covering letter (Appendix 7H, p328). The three envelopes were contained in a larger one which was addressed to the ALIS coordinator at each institution who was asked to distribute them to the appropriate head of each subject. The letters were sent out on 17th February 1997.

However, in a small number of institutions there were no students listed in some subjects, and those departments therefore received nothing. In a larger number of departments there appeared to be significant omissions from the lists of students entered for 1997 exams. At the time when the data were extracted from the database (early February 1997) not all of the 1997 exam candidates had been entered. The

numbers of departments in each of the treatment groups for whom data were available are shown in Table 41.

Table 41: Numbers of departments with data each year

TREATMENT	YEAR	ENGLISH	FRENCH	MATHS	PHYSICS	TOTAL
Departmental Information	1996	44	44	46	40	174
	1997	37	38	38	40	153
Analysis by Teacher	1996	47	42	44	43	176
	1997	40	40	37	37	154
TAMIS	1996	43	35	46	43	167
	1997	39	35	40	37	151
Control	1996	47	41	44	45	177
	1997	41	32	43	36	152
Total	1996	181	162	180	171	694
	1997	157	145	158	150	610

7.3 RESPONSES FROM INSTITUTIONS

Departmental Information

The heads of department who received the departmental analysis and target grades were invited to comment on the feedback, and just one of them responded. This was a Mathematics department whose head of department was also the ALIS and YELLIS coordinator for the institution. The gist of her reply was that, although she felt the guidance notes could have been helpful, the analysis that they had already done within the institution generally went beyond what had been sent. They were, however, quite interested in the predicted grades. Also, some students were missing from the printout sent.

Analysis by Teacher

When replies were received from departments in the 'analysis by teacher' group, the teaching group information was entered into the database and a printout for each teaching group was sent to each teacher who had taught it. A sample of this feedback, and the notes which accompanied it are provided in Appendices 7E, 7F and 7G.

Responses to the offer were received from 26 departments, of which 25 received the teacher-by-teacher analysis of their 1996 results (the other department was asked to supply some additional required information but did not reply). However, 13 of these 25 had only one teaching set in their 1996 entry, so the 'analysis by teacher' was in effect almost identical to the departmental analysis they had already received from ALIS. Moreover, of the 12 departments for whom the analysis of their 1996 results did provide new information, there were no 1997 results for four of them, either because they had no candidates in that subject that year, or because their results did not get returned and entered in the database in time for them to be included in the analysis (i.e. by Christmas 1997).

Of the eight departments who replied and had more than one teaching set and whose 1997 results were available, the vast majority of results were from four large departments who had at least ten sets in both years. Of the remaining departments, three had two sets each year and the other had four each year.

The response to the offer of teacher-by-teacher analysis was rather disappointing, and perhaps somewhat surprising. Possible explanations include that the heads of department were too busy, that they had already done such an analysis, that the teachers involved did not wish to have this information or were anxious about its becoming available to others within their institution.

TAMIS

The only comments received from anyone who had had the TAMIS disk were from the head of English in the same institution as replied to the 'Departmental Information'. His response was that that it seemed to provide information and a structure for record keeping that they already had.

Only one institution made use of the TAMIS helpline, with a single call making a fairly routine and general enquiry about the software.

7.4 EXAMINATION PERFORMANCE DATA

As can be seen from Table 41, the number of departments for whom data were available in 1997 was significantly less than the number for 1996. There are two reasons for this. Firstly, a few institutions seem to have dropped out of the ALIS project, despite apparently having been registered for 1997. Secondly, although the analysis of the 1997 results was not done until the beginning of 1998, there were still some departments whose A level results had not yet been returned and entered into the database by that time.

For those departments in the Analysis by Teacher group, the 1997 results had to be matched with the teaching sets data they had previously returned. This problem was similar to that described for the Project 1 data in Chapter 5. When the data had been matched, a further set of feedback was sent to each teacher containing their 1996 and 1997 results and a summary of the performance of all the students they had taught.

Chapter 8

Project 2: Analysis and Interpretation of Findings

The data from Project 2 consisted only of examination results. These were analysed once again using both the ALIS ordinary least squares (OLS) models and multilevel models.

Project 2 was a much larger sample than Project 1 (16,391 examination results in 1997, as compared with 504) so the reservations expressed previously about the robustness of the multilevel models used in Project 1 did not really apply to the Project 2 sample. The ability of the multilevel models to incorporate the clustering of students within departments made them very much the preferred method. However, it was thought that a brief analysis of the ALIS residuals would also be worthwhile.

8.1 OLS ANALYSIS

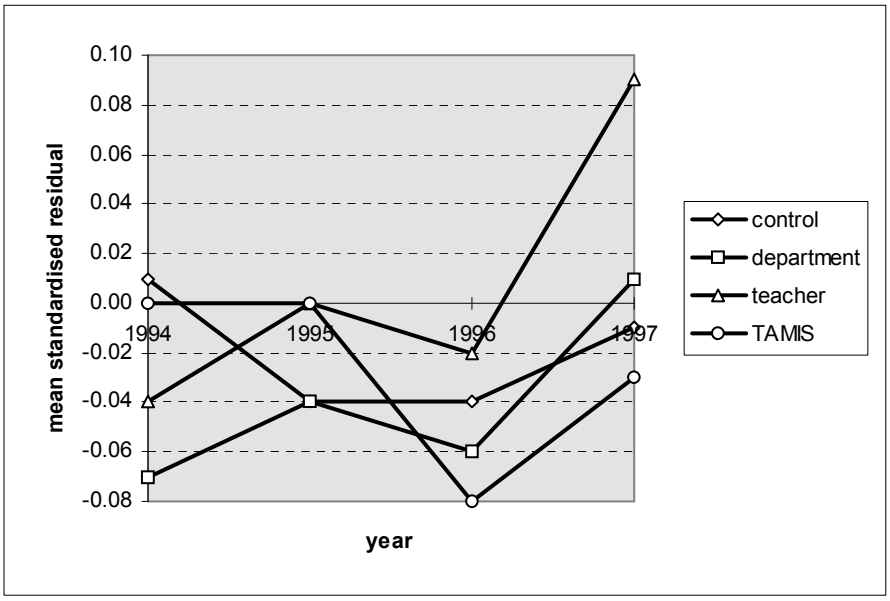
The ALIS regression equations were used to calculate the standardised residual for each student result in each of the departments in the sample for each of the years 1994 to 1997. Averages of these residuals were calculated for the different treatment groups, and for the different subjects. Of course, many of the institutions were not in the ALIS project in 1994 and 1995 so the averages for those years are not strictly comparable with those in 1996 and 1997. The total numbers of results for each treatment group in each year are shown in Table 42. The mean standardised residuals are shown graphically in Figure 12, and the values for the two years before the

experiment began are included in order to give an impression of the variability of these figures.

Table 42: Numbers of examination results in experimental sample, split by treatment

YEAR	CONTROL	DEPARTMENT INFORMATION	ANALYSIS BY TEACHER	TAMIS
1994	1630	1626	1918	1677
1995	2471	2259	2565	2491
1996	3756	3693	4113	3828
1997	4225	3853	4073	4240

Figure 12: Average residuals for each treatment group, 1994-7



The graph shows that the results of the students in the Analysis by Teacher group (shown as 'teacher') have improved the most in 1997, and it is tempting to think that this improvement is beyond the natural year-to-year variation shown by the other averages. A more systematic estimate of this variation (i.e. some kind of standard error) really requires a multilevel model, since the standard error for individual student residuals will not take account of the shared effect of being in the same department. A number of analyses using multilevel models are presented below (p224).

What is clear, however, is that the changes in performance of students in the other treatment groups are certainly not greater following the intervention than the changes in other years. In other words, there were no clear effects of either sending the 'Departmental Information' or the 'TAMIS' software on subsequent student performance.

Figure 13: Average residuals for each treatment group, 1994-7, split by subject type

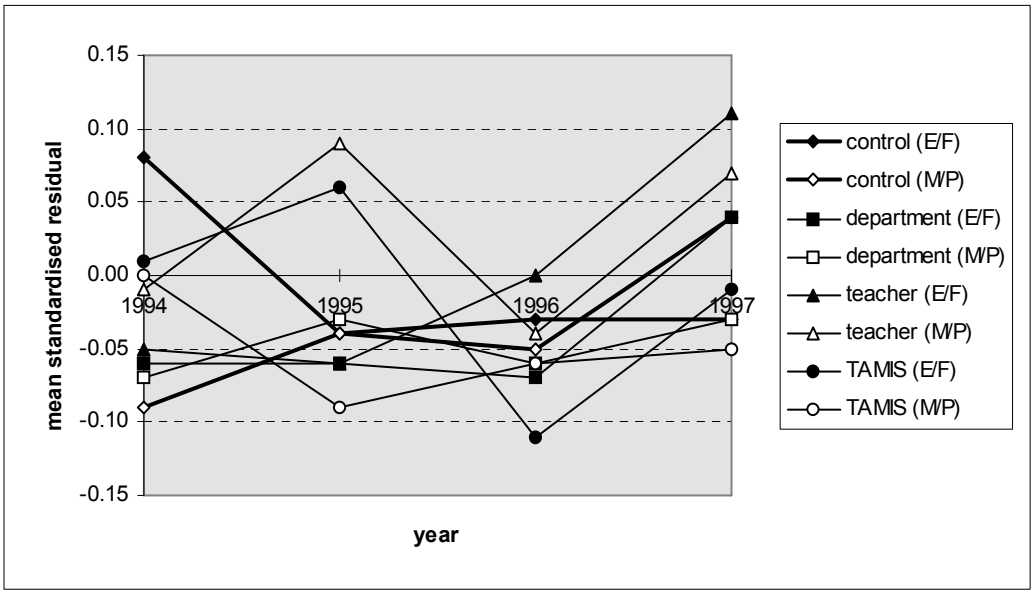


Figure 13 shows the same averages, but this time separated by subject type, with each treatment group calculated separately for English and French (E/F) and for Maths and Physics (M/P). With these smaller subgroups there is more variation. Although the 'Analysis by Teacher' groups are both still the best performers in 1997, their improvement over their 1996 performance is matched by two other subgroups ('Departmental Information' and 'TAMIS', both in English and French). From this picture it would be hard to justify claiming that any of the treatments had had a clear effect on performance.

8.2 MULTILEVEL MODELS

Model 1: Students within departments, within institutions

The first model fitted to the 1997 exam results used a three level hierarchy, nesting 14128 results within 581 departments within 157 institutions.[27] The model was fitted in stages, initially with unadjusted A level grades and then adjustment for average GCSE (coded as A*=8, A=7, B=6, etc.), parental occupation (coded as the average of both parents' scores on the Registrar General's scale with 1=unskilled, … 6=professional) and sex (coded as 1=female, 0=male). A dummy variable was used for each of the three treatments, so the coefficients of the explanatory variables may be though of as describing the relationships for the control group, and the coefficients of the treatment variables as the average difference for the results in each of those groups. Table 43 shows the parameter estimates for these models.

[27] The apparent loss of over 2000 results from the figures in Table 42 is a consequence of the inability of the multilevel modelling program *MLn* to cope with missing variables. In order to be able to use the variables 'sex' and 'parental occupation' in the analysis, all cases for which those values were not available had to be deleted from the dataset.

Table 43: Model 1: 3-level ML models for 1997 exam results

	Model 1a A level grade, unadjusted			**Model 1b** A level grade, adjusted for GCSE		
Fixed Effects Coefficients						
	estimate	**(s.e.)**		**estimate**	**(s.e.)**	
intercept	5.43	(0.14)		-9.87	(0.20)	
average GCSE				2.54	(0.03)	
parent occupation						
sex						
treatments:						
dept. info	-0.01	(0.14)		-0.01	(0.15)	
analysis by tchr	0.10	(0.16)		0.04	(0.15)	
TAMIS	-0.32	(0.15)		-0.28	(0.15)	
Random Effects Variance						
	est.	**(s.e.)**	**%**	**est.**	**(s.e.)**	**%**
between instns	1.02	(0.17)	9	0.04	(0.06)	1
between depts	0.90	(0.11)	8	1.13	(0.11)	16
between students	9.59	(0.12)	83	6.05	(0.07)	84
Goodness of Fit						
-2loglikelihood	72 742.9			66 313.9		
p(improved)				0.0000		

	Model 1c A level grade, adjusted for GCSE and parental occupation			**Model 1d** A level grade, adjusted for GCSE, parental occupation and sex		
Fixed Effects Coefficients						
	estimate	**(s.e.)**		**estimate**	**(s.e.)**	
intercept	-10.21	(0.21)		-10.22	(0.21)	
average GCSE	2.51	(0.03)		2.56	(0.03)	
parent occupation	0.11	(0.02)		0.10	(0.02)	
sex				-0.46	(0.05)	
treatments:						
dept. info	-0.01	(0.15)		-0.02	(0.16)	
analysis by tchr	0.04	(0.15)		0.02	(0.16)	
TAMIS	-0.27	(0.15)		-0.28	(0.16)	
Random Effects Variance						
	est.	**(s.e.)**	**%**	**est.**	**(s.e.)**	**%**
between instns	0.03	(0.06)	0	0.01	(0.06)	0
between depts	1.13	(0.11)	16	1.24	(0.12)	17
between students	6.04	(0.07)	84	5.99	(0.07)	83
Goodness of Fit						
-2loglikelihood	66 289.1			66 200.1		
p(improved)	6x10-7			4x10-21		

It can be seen that the incorporation of all the explanatory variables (average GCSE, parental occupation and sex) is justified statistically, in terms of the fit of the

model. Model 1d may therefore be viewed as the best fit. The coefficient of average GCSE shows that this variable is substantially the most important: a difference of one grade in incoming average GCSE score is associated with on average about a grade and a quarter difference (2.56 on the two points to a grade scale) in A level performance. This indicates that, for example, a student starting an A level course with all A*s at GCSE (i.e. a score of 8) might be expected to outperform a student with an average of Cs (i.e. a score of 5) by almost four A level grades. The model predicts on average an 'A' for the former and an 'E' for the latter. Once average GCSE grade has been included (i.e. from model 1b onwards), the total residual variance remains unchanged at about 7.2. This suggests that although the variables parental occupation and sex improve the statistical fit of the model, they do not reduce the amount of 'error' in the predictions that can be made from it. The coefficients of 0.10 and –0.46 for parental occupation and sex respectively indicate that having both parents with 'professional' occupations or being male are each associated with about a quarter grade advantage on average over those with 'unskilled' parents or who are female.

It can also be seen that almost none of the variance is between institutions, but quite a reasonable proportion is between departments (17% in model 1d). This suggests that the performances of departments within each institution were almost uncorrelated. However, as only four departments were taken from each institution (and some institutions had data for fewer than four), it is not clear how robust this result may be.

The coefficients of the treatment dummies are small (and below the level of statistical significance) in all versions of this model, suggesting that none of the treatments had any appreciable effect on students' exam performance, with the possible exception of TAMIS. The departments who received the TAMIS disk appear to have done slightly worse (the coefficient of -0.28 being equivalent to about one seventh of an A level grade) than the control group, but the difference is within the variation that could be expected by chance.

Model 2: Treatment groups subdivided

Two of the treatment groups had been subdivided, and the differences between subgroups were investigated in model 2. The dummy variable 'set info sent' was used to identify the 23 departments in the 'analysis by tchr' group who had actually responded to the offer to have set-by-set analysis, and to whom it was sent. This subdivision was therefore not on the basis of random allocation, but self selection. The TAMIS group, on the other hand, was randomly subdivided into those who were sent the offer of a telephone helpline (identified by the dummy 'TAMIS helpline') and those who were not. The coefficients of 'analysis by tchr' and 'TAMIS' in model 2 (Table 44) may therefore be interpreted as the average treatment effect for those who did not receive the analysis by teacher or the helpline offer respectively, and the coefficients of 'set info sent' and 'TAMIS helpline' represent the additional effect for those who did.

Once again, the treatment coefficients are generally small, and all below the level required to reject a null hypothesis of no effect. Curiously, the performance of those who received the set-by-set analysis seems to have been, if anything, slightly worse than that of those who did not, and, even more curiously, the offer of a telephone helpline had the biggest negative 'effect' of all. However, none of these differences are large enough to be considered either statistically or educationally significant.

Table 44: Model 2: Treatment groups subdivided

	Model 2 A level grade, adjusted for GCSE, parental occupation and sex		
Fixed Effects Coefficients			
	estimate	**(s.e.)**	
intercept	-10.22	(0.21)	
average GCSE	2.56	(0.03)	
parent occupation	0.10	(0.02)	
sex	-0.46	(0.05)	
treatments:			
dept. info	-0.02	(0.16)	
analysis by tchr	0.04	(0.16)	
TAMIS	-0.12	(0.19)	
analysis by tchr sent	-0.14	(0.32)	
TAMIS helpline	-0.34	(0.22)	
Random Effects Variance			
	est.	**(s.e.)**	**%**
between instns	0.01	(0.06)	0
between depts	1.24	(0.11)	17
between students	5.99	(0.07)	83
Goodness of Fit			
-2loglikelihood	66 197.7		

Model 3: Different subject coefficients (2 levels)

In model 3 the intercepts and slopes were allowed to vary across subjects, and the coefficients for this model are shown in Table 35, along with the regression coefficients from the OLS models used by ALIS. The model used only two levels of the hierarchy, since in a given subject there was only one department within each institution, and the negligible amount of variance accounted for by the institution in models 1 and 2 suggested that the inclusion of the third level would have very little impact on the model.

Table 45: Model 3: Different subject coefficients, 2 level ML model

		Model 3 2 levels: students, depts		**ALIS coefficients** (OLS)
Fixed Effects Coefficients				
		estimate	**(s.e.)**	
English	Intercept	-8.17	(0.25)	-8.44
	average GCSE	2.36	(0.04)	2.42
French	Intercept	-12.67	(0.43)	-14.17
	average GCSE	2.84	(0.07)	3.07
Maths	Intercept	-11.62	(0.39)	-11.35
	average GCSE	2.73	(0.06)	2.71
Physics	Intercept	-11.57	(0.48)	-12.62
	average GCSE	2.72	(0.07)	2.88
	parental occupation	0.10	(0.02)	
	sex	-0.50	(0.05)	
treatment				
	dept info	0.08	(0.15)	
	analysis by tchr	0.12	(0.13)	
	TAMIS	-0.15	(0.15)	
	set info sent	-0.32	(0.24)	
	TAMIS helpline	-0.07	(0.17)	
Random Effects Variance				
		est.	**(s.e.)**	**%**
English	depts	0.34	0.06	6
	students	4.97	0.09	94
French	depts	0.94	0.61	13
	students	6.31	0.60	87
Maths	depts	0.89	0.17	11
	students	7.37	0.17	89
Physics	depts	1.08	0.20	16
	students	5.61	0.19	84
-2logLikelihood		65 770.2		

It can be seen that the coefficients from the fixed effects part of the model are reasonably close to those estimated by ALIS, given that the ALIS models do not include parental occupation or sex. The proportions of variance within departments is generally less than in model 2. This may be because part of what appeared to be a 'departmental effect' in model 2 was due to the different relationships between average GCSE and A level grade for each different subject (i.e. different subject

difficulties): if, for example, French was a relatively hard subject, one would expect the results of students in the same department to share the influence of that disadvantage in model 2, but not in model 3.

Model 4: Adjustment for previous departmental performance

The final multilevel model used (model 4) was essentially the same as model 3, but with the inclusion of a variable for the department's residual in 1996. This residual was calculated for the 1996 data from a further multilevel model similar to model 3 but without the treatment dummies. The parameter estimates for the 1996 model are shown in Table 46, together with the regression coefficients for the ALIS OLS models for that year.

Table 46: ML model for 1996 data used to estimate departmental residuals

		1996 data 2 levels: students, depts		**ALIS coefficients** (OLS)
Fixed Effects Coefficients				
		estimate	**(s.e.)**	
English	Intercept	-8.35	(0.26)	-8.22
	average GCSE	2.40	(0.04)	2.41
French	Intercept	-15.77	(0.66)	-13.70
	average GCSE	3.33	(0.10)	3.07
Maths	Intercept	-11.65	(0.41)	-12.10
	average GCSE	2.77	(0.07)	2.88
Physics	Intercept	-13.48	(0.51)	-13.26
	average GCSE	3.03	(0.08)	3.01
	parental occupation	0.06	(0.02)	
	sex	-0.20	(0.05)	
Random Effects Variance				
		est.	**(s.e.)**	**%**
English	depts	0.61	0.10	11
	students	4.84	0.10	89
French	depts	0.86	0.19	13
	students	5.76	0.22	87
Maths	depts	0.86	0.15	10
	students	7.47	0.18	90
Physics	depts	0.73	0.16	11
	students	5.93	0.20	89
-2logLikelihood		56 812.9		

Finally, Table 47 shows the parameter estimates for model 4. All the coefficients of the treatment dummies are still below the level of statistical significance. The two treatment sub-groups have once more done worse than their respective treatment groups as a whole, but again the difference is not enough to make its attribution to chance seem implausible. Hence we must once again conclude that there were no clear effects of any of the experimental treatments in Project 2.

Table 47: Model 4: Adjustment for previous departmental performance

		Model 4 A level grade, adjusted for GCSE, parental occupation, sex and dept's previous residual		
Fixed Effects Coefficients				
		estimate	**(s.e.)**	
English	Intercept	-8.07	(0.27)	
	average GCSE	2.35	(0.04)	
French	Intercept	-12.46	(0.45)	
	average GCSE	2.82	(0.07)	
Maths	Intercept	-11.45	(0.41)	
	average GCSE	2.70	(0.06)	
Physics	Intercept	-10.90	(0.51)	
	average GCSE	2.64	(0.08)	
	parental occupation	0.08	(0.02)	
	sex	-0.50	(0.05)	
	previous dept residual	0.51	(0.07)	
treatment				
	dept info	0.10	(0.13)	
	analysis by tchr	0.11	(0.13)	
	TAMIS	0.04	(0.15)	
	set info sent	-0.24	(0.23)	
	TAMIS helpline	-0.30	(0.18)	
Random Effects Variance				
		est.	**(s.e.)**	**%**
English	depts	0.37	0.07	7
	students	5.03	0.10	93
French	depts	0.71	0.17	10
	students	6.37	0.19	90
Maths	depts	0.57	0.12	7
	students	7.38	0.18	93
Physics	depts	0.84	0.18	13
	students	5.56	0.20	87
-2logLikelihood		60 033.7		

It is interesting to note that the regression coefficient for the previous year's residual is 0.51, compared with the 0.35 found in model 4 for Project 1 (see p202). As the Project 2 sample was larger and more representative it will almost certainly have provided the more accurate estimate. Hence we should modify the finding

reported in Chapter 6 to say that students in a department with a good previous performance might be expected to benefit by approximately half the previous residual, rather than the one third estimated before. Students in a department whose results last year were a grade above the average expectation could be expected to be about half a grade better this year. It therefore seems that, particularly in departments with extreme residuals, knowledge of a department's results last year does make a difference to the prediction one would make for an individual student this year. However, the inclusion of this variable produces only a negligible decrease in residual variances: the overall accuracy of the individual predictions that can be made is scarcely changed.

Chapter 9

Summary and Discussion

9.1 SUMMARY OF RESULTS

The main findings of the empirical study are summarised below.

Project 1

- Project 1 was an experiment involving 9 volunteer institutions, all of which had been members of ALIS for at least three years.
- Exploratory interviews and a pilot questionnaire were used to develop a questionnaire instrument. This was designed to measure teachers' attitudes, self-perceptions and self-reported behaviour relevant to their uses of feedback.
- Questionnaires were sent to 157 teachers of A level English, French, Mathematics and Physics. 73 (47%) were returned. Institutions varied widely in their rates of response. Response rates for teachers across subjects were broadly equal.
- Institutions were asked to identify students' teaching sets. Delays or non-response reduced the number of teachers available for the experiment to 44, all but 9 of whom were drawn from just 3 institutions.
- Departments were paired by size within each institution and randomly allocated to feedback or control groups. 22 teachers were allocated to each.

- Teachers in the feedback group were sent information about the value added performance and attitudes of students in each of the classes they had taught over the last three years.
- Teachers varied widely in their reported ease of understanding of the feedback. In particular, teachers of numeric subjects (Mathematics and Physics) reported significantly less difficulty than teachers of French and English.
- A modified version of the questionnaire was used after the feedback to measure attitude changes. Nine attitude constructs were derived, most of which had adequate reliability (test-retest and internal consistency). However, a comparison between attitudes inferred from open ended comments and the corresponding attitude constructs showed only moderate correlation.
- The largest apparent effects of the feedback on attitude changes were in attitudes towards ALIS, which seemed to have become more negative for those who received the feedback. Quite substantial effect sizes of 0.5-0.6 were found, but these did not quite achieve statistical significance ($p \geq 0.05$). The negative effects of the feedback on attitudes towards ALIS appeared to be greater for teachers of non-numeric subjects (French and English).
- A large majority of the teachers said that feedback based on class by class analysis of performance should be sent to the head of department and to the institution's ALIS coordinator. Just over half those who expressed a view said that it should be sent to the class teacher. There were no differences in these views between those who had received the feedback and those who had not.
- Interviews conducted with six of the teachers suggested that, for some of them at least, their scores on the questionnaire constructs did not correspond with their own perceptions, and in particular with their perceptions of any changes.
- Examination performance and attitudes of students taught by teachers in the two groups were compared. The former was analysed using raw results, ALIS residuals and multilevel models.
- The timing of the measurement of student attitudes made it very unlikely that a feedback effect would be found, and the analysis was inconclusive.

- Students in feedback and control groups were well matched in terms of prior attainment (GCSE scores) and parental occupations. A level grades in the same departments prior to the experiment were also well matched. After the experiment, students of the teachers who had received the feedback achieved about half an A level grade better than those in the control group.
- In terms of ALIS residuals (i.e. after adjusting for prior attainment), students in the feedback group outperformed those in the control by about a third of a grade (effect size 0.3).
- A number of multilevel models were used. These all gave estimates (adjusted for parental occupation and prior achievement, and taking into account the effects of 'clustering' in departments) of the feedback effect at around one third of an A level grade (effect sizes around 0.2-0.3).
- Teachers' 'performance' (as measured by the average of their students' ALIS residuals each year) was found to be only moderately stable. When averages where restricted to five or more students, a correlation between successive years of 0.6 was found.
- With the teacher as the unit of analysis (using average students' performance), the apparent effects of the feedback were quite substantial. Estimates of the effect size were between 0.3 and 0.7. However, the sample was too small for these to achieve statistical significance ($p \geq 0.05$) and substantial sample attrition made causal inference somewhat problematic. Apparent effect sizes were even larger (0.7-1.4) for teachers whose previous performance was below average and (0.7-1.5) for the teachers who said class by class feedback should be sent to the class teacher. Inferences from these subgroups, however, are even less secure.

Project 2

- Project 2 was an experiment involving 192 institutions, randomly selected from the membership of the ALIS project in 1996-7.
- Four departments (English, French, Mathematics and Physics) in each institution were randomly allocated to receive either analysis and target grades for all the

students in their department, to be offered a class by class analysis if they returned a class list, to be sent a piece of software for DIY analysis or to be used as the control.

- A level examination results of all the departments were analysed using ALIS residuals and multilevel models. No significant differences were found between any of the treatment groups.
- Multilevel models showed average GCSE score to be substantially the best 'predictor' of A level grade, although the inclusion of parental occupation and sex did improve the statistical fit of the model. A model incorporating the department's residual from last year as an explanatory variable estimated its regression coefficient as 0.5.
- Multilevel models estimated the proportion of variance in adjusted outcome between institutions to be close to zero, and the proportion between departments at about 17%. When different subjects were modelled separately, the proportion between departments varied from 6 to 16%.

9.2 DISCUSSION

The results of the empirical investigation have been reported in detail in chapters 6 and 8, and summarised above. Many of the implications of the findings have already been discussed, so the following serves mainly to draw out some recurring themes and to set the conclusions in the context of other research.

Security of inferences

A number of threats have been identified to the security of any inferences that can be drawn from the study. These have been largely concerned with two issues in Project 1: the representativeness of the sample and the validity of the attitude constructs.

Representativeness of the sample

A number of factors have contributed to this concern, in particular the fact that the participants were volunteers, the high rate of attrition of the sample and its resulting small size.

The volunteer status of the teachers in the study, along with their relative expertise in the use of ALIS data, certainly makes it arguable that none of the findings would transfer to the wider population of teachers and schools. With the benefit of hindsight, it seems likely that the level of teachers' familiarity with the kinds of feedback provided may be quite important. Indeed, this factor may account for the difference between Project 1 and Project 2 in the effects of the feedback on student performance. On the other hand, some of the teachers in Project 1 were not at all familiar with ALIS feedback and its effect on them was, if anything, greater. The issue is complex and cannot really be said to be well understood.

Validity of the attitude constructs

One of the main things that I have learnt from doing this research is that attitudes are hard to measure. Despite following accepted practice in the design of the questionnaire instrument, and despite the resulting attitude constructs having achieved acceptable standards of reliability, the attempts to validate them against peoples' self-perceptions were rather disappointing. Evidently, measuring attitudes in a meaningful way is far from straightforward.

Not being able to interpret confidently the attitudes of the people in this study is something of a handicap. Feedback *per se* can have no effect on anything; it is only through the processing of the feedback that any effects will be realised. Hence it seems quite important to know how the feedback was received and processed if we are to understand how it had its effects.

Need for replication

The point has already been made that the findings from any single study may be too dependent on the particular sample and methodology used to be a safe basis for general conclusions. This is a general issue and the findings from this study, like

those of any other, would become substantially more secure if they were to be replicated.

On a more personal level, I feel that having completed the study, I now have a much better understanding of what the important variables are, how they should be operationalised and which methodologies are appropriate. Only now that I have finished am I really ready to begin to do the research properly.

Models of school effectiveness

Despite the critical approach adopted in Chapter 2, some of the analysis of this study has lapsed into the very practices that were held up as indefensible. Examples include the use of average residuals as a measure of teachers' performance and the use of 'non-explanatory' explanatory variables such as sex and parental occupation in the modelling. Two issues now seem particularly important for the future of school effectiveness research.

Firstly, the question of stability seems crucial. If the same teacher, or school, does not produce consistent measures of 'effectiveness' across different classes or years, then we really cannot claim to be measuring effectiveness at all – or at best measuring it only inaccurately. It may be that value added can successfully measure student progress, but to equate that with teacher effectiveness would be poor modelling, not to mention unjust.

Secondly, the statistical modelling employed in school effectiveness research badly needs to be guided by some understanding. In particular, the match between teachers' objectives and the outcomes measured needs considerably more attention. Also in need of attention are the processes and levels at which effects should be sought. Finally, the use of explanatory variables that really do explain would be a major step forward. Perhaps when we start to understand sex differences in terms of thinking styles, values or differences in maturation, or can explain the 'effect' of SES in terms of factors such as differing aspirations, expectations, resources or cultural dissonance will we begin to produce models of school effectiveness that contribute to improvement.

Feedback as a means of school improvement

My original feeling that feedback could be a promising way of enhancing performance has not been diminished by conducting this study. Indeed, finding an effect of the order of one third of an A level grade in Project 1 has given it some encouragement. However, the main finding in both the existing literature and in this research seems to be that the effects of feedback are extremely complicated: sometimes large effects are found; sometimes they are negative. When it is remembered that the teachers in the control group were also receiving feedback from ALIS, and many people in both groups seemed to be already doing the kinds of additional analysis that I sent them, the effects found seem surprisingly large.

My feeling now is that the main effect of the feedback in this case was probably in focusing attention on the outcomes being measured, rather than any diagnostic function. However, a more sophisticated experimental design would be needed to test this conjecture.

Target setting: theory into practice

One of the most convincing parts of the literature on feedback effects found in the review in Chapter 3 was the theory of goal setting (Locke and Latham, 1990). This theory is arguably as well specified and comprehensively tested as any in the field of social science. It makes clear predictions, defining the conditions under which goal setting will have optimal effects on performance, and is supported by evidence unmatched in both quantity and diversity. Target setting is also an important part of the UK government's strategy to raise standards, in which LEAs and schools are required to set ambitious targets.

It was somewhat disappointing, therefore, to find that the intervention in Project 2 that provided target grades for all students in the department appeared to have no effect at all on subsequent performance, and that the 'target grades' part of the feedback sent in Project 1 received the least favourable ratings. Given the amount of investment in this particular strategy for improvement, it would be of some interest to understand better the conditions under which setting targets in schools may be expected to lead to higher achievement.

Appendices for Chapter 5:

Project 1: Data Collection

Introductory comments:

I would very much like to ask you some questions as part of my research for a PhD. I want to try to find out how teachers use the feedback that they get about their own performance, and how they feel about it.

I want to get your views and feelings, so I will ask fairly open questions. If you want me to explain more what I am getting at, then please ask for clarification.

It will help me very much if I can record our conversation. I can assure you that it will be used only for research and will be confidential between us. Are you happy for me to record it?

I would like to take ten minutes of your time now. At the end of that time I will try to draw it to a close, unless you wish to continue.

Possible questions and starting points to use, together with suitable follow-ups:

- What kinds of feedback do you get about your own teaching? (Prompt if necessary: from formal (appraisal, exam results) to informal (comments of colleagues, students, parents); from immediate (facial expressions in lesson) to much later ('value added' residuals))
- What other sources of information do you have about the quality of your work? How do you know how good a teacher you are?
- How much credibility do you give to each source of information?
- How important is feedback to you? Does it affect your view of your own performance?
- How does feedback affect your attitudes and feelings (eg encouraging, motivating, frustrating, etc, ...)
- What kinds of feedback would you like to get?

Durham University/ALIS/YELLIS Feedback Project

Questionnaire

Autumn 1996

A: Personal details:

Name: .....

School/ College: .....

Age range: .....

Position: .....

How long have you worked there? .....

Subjects and levels/ages taught: .....

B: Use of and attitudes to feedback:

Please list any forms of feedback (formal or informal) or information you have had about your performance in this job:

Please rate the following statements on a five point scale from 'agree strongly' to 'disagree strongly':

	agree strongly				disagree strongly
1. I like to receive objective feedback about the quality of my work.	❑	❑	❑	❑	❑
2. I am always keen to have my performance assessed.	❑	❑	❑	❑	❑
3. I know when I've got things right: no-one needs to tell me.	❑	❑	❑	❑	❑
4. I believe I am a good teacher.	❑	❑	❑	❑	❑
5. I do not like situations in which I am being judged.	❑	❑	❑	❑	❑
6. Being good at my job is important to me.	❑	❑	❑	❑	❑
7. If ALIS/YELLIS gave me information about my teaching performance I would find it useful and informative.	❑	❑	❑	❑	❑
8. My effectiveness as a teacher depends largely on how hard I try.	❑	❑	❑	❑	❑
9. Receiving feedback can help me to improve what I am doing.	❑	❑	❑	❑	❑

	agree strongly				disagree strongly
10. If I am not successful at some aspect of my work it is usually because the task is too hard	❑	❑	❑	❑	❑
11. The ALIS/YELLIS data on attitudes do not tell us anything worthwhile.	❑	❑	❑	❑	❑
12. I am my own sternest critic.	❑	❑	❑	❑	❑
13. I am largely responsible for the exam performance of my students.	❑	❑	❑	❑	❑
14. The residuals calculated by ALIS/YELLIS are a fair way of measuring how well students have done.	❑	❑	❑	❑	❑
15. I am worried that feedback about my teaching performance could be used against me.	❑	❑	❑	❑	❑
16. I often have doubts about whether I am doing a good job.	❑	❑	❑	❑	❑
17. If the students I teach perform badly, it is their fault.	❑	❑	❑	❑	❑
18. The quality of my teaching is reflected in the exam success of my students.	❑	❑	❑	❑	❑
19. I usually seek information with which to judge whether I am achieving what I want to.	❑	❑	❑	❑	❑
20. The exam results of the students I teach reflect my ability as a teacher.	❑	❑	❑	❑	❑
21. ALIS/YELLIS residuals do not really mean very much.	❑	❑	❑	❑	❑
22. I am concerned that information form ALIS/YELLIS could be used to check up on me.	❑	❑	❑	❑	❑
23. I feel anxious when I am evaluated.	❑	❑	❑	❑	❑
24. Whether my students do well or not depends more on them than it does on me.	❑	❑	❑	❑	❑
25. There are too many errors in the feedback provided by ALIS/YELLIS for their findings to be reliable.	❑	❑	❑	❑	❑
26. I like to receive feedback about the quality of my work.	❑	❑	❑	❑	❑
27. If the analysis by ALIS/YELLIS shows that a particular department has a high score, then I will believe that there has been some good teaching in that department.	❑	❑	❑	❑	❑
28. The students I teach get good exam grades, given their ability.	❑	❑	❑	❑	❑
29. I feel confident about the quality of my work.	❑	❑	❑	❑	❑
30. If ALIS/YELLIS gave me information about my teaching performance I would find it quite threatening.	❑	❑	❑	❑	❑
31. Doing well is more important to me when I am being assessed.	❑	❑	❑	❑	❑
32. When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough.	❑	❑	❑	❑	❑
33. I think the Head/Principal should not use ALIS/YELLIS results in staff appraisal.	❑	❑	❑	❑	❑

C: About ALIS/YELLIS:

1. In which project(s) is your institution? ALIS ❑ YELLIS ❑

2. Approximately when did you first become aware of the existence of ALIS or YELLIS?

- ❑ within the last year
- ❑ between one and three years ago
- ❑ more than three years ago

3. What information have you had from ALIS/YELLIS about the performance or attitudes of your students?

4. What use have you made of it?

5. How valuable have you found it?

6. Please tick any of the following which describe(s) the stage you are at with using ALIS/YELLIS:

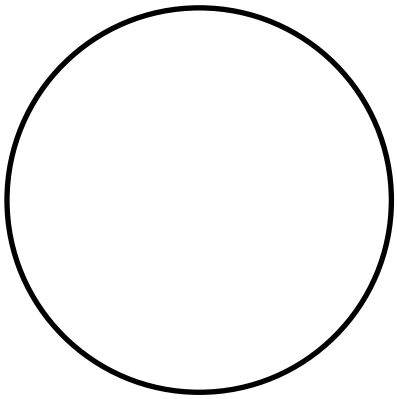
- ❑ I have not had any contact with it
- ❑ I have begun to learn about it
- ❑ I have made some plans to use it
- ❑ I have used it in ways that have been required of me
- ❑ I have made my own routine use of it
- ❑ I have applied it effectively to solve problems
- ❑ I have integrated its use into my work
- ❑ I have evaluated and modified it to meet my needs

D: Responsibility for students' exam performance:

Imagine that the circle represents the exam performance of typical students. Please divide it into sections (like a pie chart) where the size of each piece indicates the **relative importance** of that factor in determining exam performance.

Factors which affect exam performance:

A: students' ability
B: students' home background
C: students' character attributes (eg, hard working/lazy)
D: which teacher they have
E: which school they go to
F: other factors (if any particular ones, please list them)



E: Further comments:

Please make here any additional comments, including comments on any of the above questions that you found to be unclear, meaningless or otherwise hard to answer:

F: Consent to telephone:

I may find it useful to 'phone you with further questions, or to follow up something you have said. Would you be happy for me to do this? Yes ❑ No ❑

If yes, telephone no: .....

Times/day(s) when it is best to phone: .....

Thank you very much for completing this questionnaire.
Please return it in the enclosed s.a.e.

Robert Coe
Durham University School of Education
Leazes Road
Durham DH1 1TA

Durham University/ALIS Feedback Project

Questionnaire

Autumn 1996

A: Personal details

1. Name: ..... 2. Sex: ❑ M ❑ F

3. School/
College: .....

4. Position: .....

5. How long have you worked there? .....

6. Which of the following is your main subject taught at A level?

❑ English ❑ French ❑ Maths ❑ Physics

7. Do you teach any classes which will take A level in this subject this year ('97) ❑ Yes ❑ No

8. Did you teach any classes last year which took A level in this subject in '96? ❑ Yes ❑ No

B: Use of and attitudes to feedback:

Please list any forms of feedback (formal or informal) or information you have had about your performance in this job:

Please rate the following statements on a five point scale from 'agree strongly' to 'disagree strongly':

	agree strongly				disagree strongly
1. I like to receive objective feedback about the quality of my work.	❑	❑	❑	❑	❑
2. I am always keen to have my performance assessed.	❑	❑	❑	❑	❑
3. The exam results of the students I teach reflect my ability as a teacher.	❑	❑	❑	❑	❑
4. I believe I am a good teacher.	❑	❑	❑	❑	❑
5. I do not like situations in which I am being judged.	❑	❑	❑	❑	❑
6. If ALIS gave me information about my teaching performance I would find it useful and informative.	❑	❑	❑	❑	❑

	agree strongly				disagree strongly
7. My effectiveness as a teacher depends on how I choose to teach.	❑	❑	❑	❑	❑
8. Receiving feedback can help me to improve what I am doing.	❑	❑	❑	❑	❑
9. If a student who does not much like my subject joins my class, I can usually help him or her to enjoy it more.	❑	❑	❑	❑	❑
10. The ALIS data on attitudes do not tell us anything worthwhile.	❑	❑	❑	❑	❑
11. I prefer tasks in which I can see how well I am doing.	❑	❑	❑	❑	❑
12. I am responsible for the exam performance of my students.	❑	❑	❑	❑	❑
13. The value-added scores (residuals) calculated by ALIS are a fair way of measuring how well students have done.	❑	❑	❑	❑	❑
14. I am worried that feedback about my teaching performance could be used against me.	❑	❑	❑	❑	❑
15. I often have doubts about whether I am doing a good job.	❑	❑	❑	❑	❑
16. If the students I teach perform badly, it is their fault.	❑	❑	❑	❑	❑
17. The quality of my teaching is reflected in the exam success of my students.	❑	❑	❑	❑	❑
18. I usually seek information with which to judge whether I am achieving what I want to.	❑	❑	❑	❑	❑
19. My institution gets very little benefit from being in ALIS.	❑	❑	❑	❑	❑
20. I am concerned that information from ALIS could be used to check up on me.	❑	❑	❑	❑	❑
21. I feel anxious when I am evaluated.	❑	❑	❑	❑	❑
22. The A level grades that students get depend on who teaches them.	❑	❑	❑	❑	❑
23. There are too many errors in the feedback provided by ALIS for their findings to be reliable.	❑	❑	❑	❑	❑
24. If the analysis by ALIS shows that a particular department has a high score, then I will believe that there has been some good teaching in that department.	❑	❑	❑	❑	❑
25. I feel confident about the quality of my work.	❑	❑	❑	❑	❑
26. If ALIS gave me information about my teaching performance I would find it quite threatening.	❑	❑	❑	❑	❑
27. Doing well is more important to me when I am being assessed.	❑	❑	❑	❑	❑
28. When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough.	❑	❑	❑	❑	❑
29. I think the Head/Principal should not use ALIS results in staff appraisal.	❑	❑	❑	❑	❑

C: About ALIS:

1. Approximately when did you first become aware of the existence of ALIS?

- ❑ within the last year
- ❑ between one and three years ago
- ❑ more than three years ago

2. What information have you had from ALIS about the performance or attitudes of your students?

3. What use have you made of it?

4. How valuable have you found it?

5. Please tick any of the following which describe(s) the stage you are at with using ALIS:

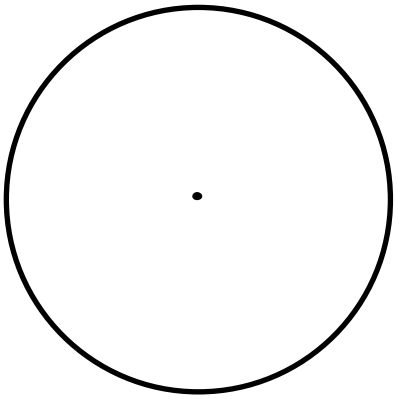
- ❑ I have not had any contact with it
- ❑ I have begun to learn about it
- ❑ I have made some plans to use it
- ❑ I have used it in ways that have been required of me
- ❑ I have made my own routine use of it
- ❑ I have applied it effectively to solve problems
- ❑ I have integrated its use into my work
- ❑ I have evaluated and modified it to meet my needs

D: Responsibility for students' exam performance:

Imagine that the circle represents the exam performance of typical students. Please divide it into sections (like a pie chart) where the size of each piece indicates the **relative importance** of that factor in determining exam performance.

Factors which affect exam performance:

A: students' ability
B: students' home background
C: students' character attributes (eg, hard working/lazy)
D: which teacher they have
E: which school/college they go to
F: other factors (if any particular ones, please list them)



E: Further comments:

Please make here any additional comments, including comments on any of the above questions that you found to be unclear, meaningless or otherwise hard to answer:

F: Consent to telephone:

I may find it useful to 'phone you with further questions, or to follow up something you have said. Would you be happy for me to do this? ❑ Yes ❑ No

If yes, telephone no: .....

Times/day(s) when it is best to phone: .....

Thank you very much for completing this questionnaire.
Please return it in the enclosed s.a.e.

Robert Coe
Durham University School of Education
Leazes Road
Durham DH1 1TA

RESULTS Page 1
Inst: 999
Subject: Maths
Exam year: 96
Set: 4

unique set ID	surname	forename	teachers	avge GCSE score	A level grade	ALIS 'pred' grade	resid	std resid	VA cat	no of subjs taken	curr bal	av std resid (all subjs)	rel VA	tot UCAS pts
999269604	XXXXXX	XXXXX	TXX RXX	5.89	6	4.86	1.14	.39	0	4	-.3	-.02	0	20
	XXXXX	XXXXXX	TXX RXX	6.70	4	7.20	-3.20	-1.09	-	3	-1.0	-.37	-	18
	XXXXXX	XXXX	TXX RXX	6.70	6	7.20	-1.20	-.41	0	4	-.5	.40	-	32
	XXXXXX	XXXX	TXX RXX	7.10	10	8.35	1.65	.57	0	4	-1.0	.65	0	40
	XXXXXX	XXXXX	TXX RXX	6.55	-2	6.75	-8.75	-3.00	-	4	.0	-.75	-	20
	XXXXXXXXX	XXXXXX	TXX RXX	4.78	2	1.66	.34	.12	0	4	-.3	1.12	-	20
	XXXXXX	XXXXXXX	TXX RXX	5.80	-2	4.60	-6.60	-2.26	-	4	-.3	-1.21	-	6
	XXXXXXX	XXXXXXXX	TXX RXX	6.67	4	7.10	-3.10	-1.06	-	5	-.6	-.35	-	30
	XXXXX	XXXXXX	TXX RXX	5.78	2	4.54	-2.54	-.87	-	4	-.5	-.37	-	14
	XXXXXX	XXXXXXX	TXX RXX	6.30	4	6.04	-2.04	-.70	-	4	-.3	.33	-	28
	XXXXXXXXX	XXXXXX	TXX RXX	5.80	-2	4.60	-6.60	-2.26	-	4	-.3	-.01	-	14
	XXXXXXX	XXXXXXX	TXX RXX	5.50	0	3.74	-3.74	-1.28	-	4	-.3	-.66	-	10
	XXXXXXXX	XXXXX	TXX RXX	7.20	10	8.64	1.36	.47	0	4	-.3	-.10	+	34
	XXXXXXX	XXXXXXXXX	TXX RXX	6.30	2	6.04	-4.04	-1.38	-	4	-.5	-.73	-	16
Mean				6.22	3	5.81	-2.67	-.91		4	-.4	-.15		22
N				14	14	14	14	14		14	14	14		14

avge GCSE score = average of all GCSEs taken (A*=8,A=7,B=6,C=5,etc)
A level grade = grade achieved, coded as A=10,B=8,C=6,D=4,E=2,N=0,U=-2
ALIS 'pred' grade = avg grade achieved by students with same GCSE score
resid = actual grade - 'pred' grade (= VALUE ADDED)
std resid = resid scaled down to enable stat comparisons across subjs & years
VA cat = value added category, classified as '+' (top 25%), '0' (middle 50%), '-' (bottom 25%)
no of subjs taken = includes all with resid calculated (A & AS level + GNVQ)
curr bal = curriculum balance, coded -1 (all science) to 1 (all arts)
av std resid (all subjs) = value added avg for all A levels taken
rel VA = value added in your subj relative to other subjs, classified as '+' (top 25%), '0' (middle 50%), '-' (bottom 25%)
tot UCAS pts = total points for all grades achieved (A=10,B=8,C=6,D=4,E=2,N=0,U=-2, AS grades count half)

SUMMARY BY TEACHER

This table summarises the value added performance of all the sets you have taught

institution	subject	teacher	exam year	teaching set	teachers	no of students	mean std resid	sig level	mean resids in all subjs	proportion of teaching of set	equivalent no of students (weight)
999	Maths	RXX	94	3	RXX	6	.27	.51	.43	1.00	6.00
				7	RXX	12	.00	1.00	-.03	1.00	12.00
				10	RXX CXX	11	-.05	.87	-.08	.50	5.50
			95	1	PXX DXX RXX	1	.71	.48	.61	.33	.33
			96	4	TXX RXX	14	-.91	.00	-.15	.50	7.00
				10	FXX TXX RXX	8	.37	.30	.38	.25	2.00
		summary	.			52	-.12	.48	.05	.	32.83

mean std resid = average value added performance of all students in that set
sig level = probability that a mean value as large (or small) as this would happen purely by chance (ie nothing to do with tchng)
mean resids in all subjs = avg value added perf of students in this set across all their subjects (compare this with mean std resid)
proportion of teaching of set = propn taught by you
equivalent no of students (weight) = [no of stds] x [propn taught by you]

summary = totals/weighted averages of each variable across all sets

CLASS AVERAGES

Instn: 999
Subj: Maths

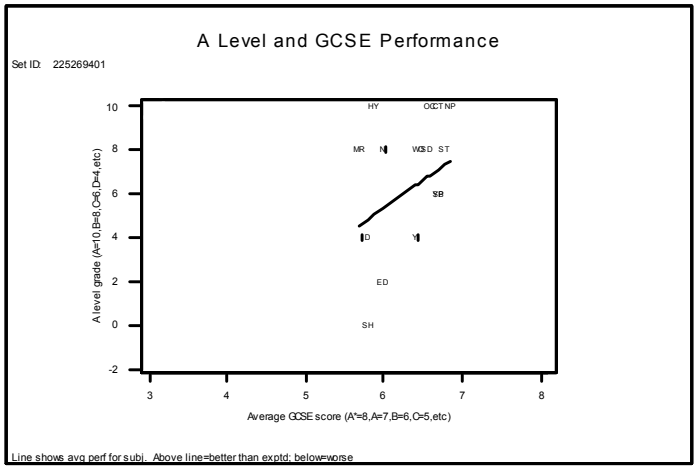
This table shows average values on a range of variables collected by ALIS
for each of the classes you have taught, compared with the average for all students
in your department that year and all in your subject that year in the ALIS cohort.

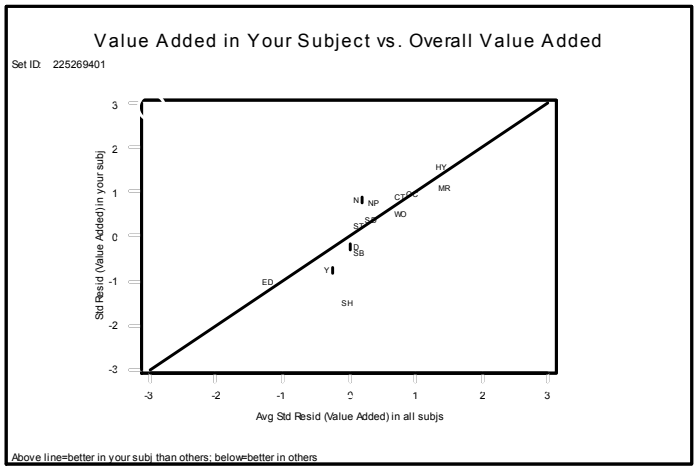
dept ID	teacher	exam year	avge for:	tchng set	no of studnts	avge GCSE score	ITDA	parent occupn	% female	liklhd staying in edn	A level grade	av std resid	res (all subjs)	att to subj
99926	RXX	94	your classes:	3	6	6.19	.	4.50	17	.	6.67	.27	.43	.
				7	12	6.44	.	4.79	33	.	7.33	.00	-.03	.
				10	12	5.52	.	4.13	25	.	4.33	-.05	-.08	.
			your dept:	.	95	6.18	.	4.60	24	.	6.32	.03	-.02	.
			your subj:	.	9883	6.03	.	4.36	34	4.14	5.43	.00	.00	3.34
95			your classes:	1	10	6.76	.	4.80	50	4.29	10.00	.71	.61	3.80
				9	6	5.74	.	4.75	0	4.29	.	.	.30	1.95
			your dept:	.	102	6.05	.	4.79	38	4.13	5.60	-.07	.08	3.14
			your subj:	.	15835	5.97	61.80	4.37	35	4.16	5.74	.00	.00	3.25
96			your classes:	4	14	6.22	.	3.18	14	4.29	3.14	-.91	-.15	.
				10	10	5.99	.	5.00	60	4.03	6.75	.37	.38	2.57
			your dept:	.	141	6.24	.	3.23	38	4.20	5.20	-.25	.12	2.57
			your subj:	.	33387	6.16	59.90	4.44	35	4.15	6.13	.00	.00	3.35

avge GCSE score = avge of all GCSEs taken (A*=8,A=7,B=6,C=5,etc)
ITDA = International Test of Developed Abilities, a test of general academic ability provided by ALIS
parent occupn = parents' occupations from 1 (unskilled) to 6 (professional)
% female = percentage of students in that group who were female
liklhd staying in edn = likelihood of continuing to HE, training etc, 1=low, 5=high
A level grade = grade achieved, coded as A=10,B=8,C=6,D=4,E=2,N=0,U=-2
std resid = a measure of value added performance in your subject
av res (all subjs) = a measure of value added performance of those students in all their subjects
att to subj = questnre responses on attitude to your subject, from 1 (negative) to 5 (positive)

Performance of students in this set

(Each student identified by initials; position shows performance)





DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT
Durham University School of Education, Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message);
Fax: 0191 374 3506; Email: r.j.coe@durham.ac.uk

Suggestions for using this information:

1. Identify students who have performed significantly better (or worse) than would have been expected from their prior attainment.

(Using information from the 'RESULTS' printout and 'Performance of students in this set' graphs)

(a). Identify these individuals:

- From the first graph ('A level and GCSE Performance'): Students whose initials are plotted above the line have achieved a grade which is better than the average of those with similar GCSE grades (ie have positive residuals). Any which look to be a long way above (or below) the line and the rest of the group may be identified as 'over-(under-)achievers'.
- From their residuals (in the 'resid' column of the 'RESULTS' printout): A residual of +2 means they achieved one grade above expectation. Residuals above +4 (or less than -4) indicate performance more than two grades away from what might have been expected. This is a significant difference and these students are worth further attention.
- From the value added category ('VA cat'): This classification is based on the frequency of occurrence of large residuals. Any students classified as '+' or '-' are in the top 25% or the bottom 25% respectively and may therefore be seen as having 'over-(or under-)performed'.

(b). For each of these students:

- Is it fair to describe them as having 'over-performed' ('under-performed')?
- If so, can you account for their performance? (See also 2(b), below.)
- Can anything be learnt that might benefit current or future students?

©. Consider the students in each group collectively:

- Are there any common features among them?
- What proportion of your students are in each of the '+' and '-' categories (compared with the expected quarter in an 'average' group)?
- Are any particular subgroups (eg males/females) over- or under-represented in either group? (One way to look for this is to colour code the initials on the graph.)
- Are there any patterns in the spread of initials on the graph? (Eg those with lower GCSEs mostly below the line, higher GCSEs mostly above would suggest that the more able are doing better in value added terms.)

2. Compare students' value added performance in your subject with that in their other subjects.

(Using information from the 'RESULTS' printout and 'Performance of students in this set' graphs)

(a). Compare the two for each student:

- From the second graph ('Value Added in Your Subject vs. Overall Value Added'): Students whose initials are plotted above the line have performed better (in terms of value added) in your subject than in their other subjects.
- From the 'rel VA' category, which compares their standardised residuals in your subject ('std resid') with their average for all subjects ('av std resid (all subjs)'). Those whose value added performance in your subject is better than that in their other subjects are coded '+'; '0' indicates the two were about the same; '-' shows they did better in their other subjects.

(b). If the two values are similar (ie 'rel VA' is coded '0' & initials are plotted close to the line):

- If you previously identified this student as having over- (or under-) achieved (1(a), above), it may be that any credit (or blame!) for an apparently good (or poor) performance is not due to you - since they have done equally well in all their other subjects. Among possible explanations are that their GCSE grades were not a true reflection of their ability, that they worked particularly hard in all subjects, that they suffered some personal event which affected their studies, etc, etc …

©. If the two values are significantly different (ie 'rel VA' is coded '+' or '-' & initials are plotted away from the line):

- Is it fair to describe them as having 'over-performed' ('under-performed') relative to their other subjects?
- If so, can you account for their performance?
- Can anything be learnt that might benefit current or future students?
- Are there any common features among each group? (Eg particular subject combinations.)
- Are any particular subgroups (eg males/females) over- or under-represented in either group? (Again, colour code the initials on the graph.)

3. Compare characteristics of each class taught by you with those of all students in your department and the whole ALIS cohort.

(Using information from the 'CLASS AVERAGES' printout)

(a). Compare the intake characteristics:

- *General academic ability*, as measured by average GCSE score (the best predictor of A level performance) and the ITDA (International Test of Developed Abilities: this test is provided free by ALIS, but not all institutions use it).
- *Socio-economic status*, as measured by parents' occupations.
- *Gender balance* (% female).
- *Aspirations*, as measured by the likelihood of staying in education (LSE) scale. This is produced from responses to the questionnaire students complete at the beginning of the A level course.

(b). Compare the measured outcomes:

- *A level grades*.
- *Value added* ('std resid').
- *Value added, relative to value added in other subjects* (compare 'std resid' with 'av res (all subjs)').
- *Attitudes* to your subject.

4. Interpret the statistical significance of the value added performance of each class and of all the students collectively.

(Using information from the 'SUMMARY BY TEACHER' printout)

The 'sig level' for each teaching set shows the probability that a randomly selected group of 'average' students would get a value added average as extreme as this. It depends on the size of the 'mean std resid' and the number of students in the group. Statisticians conventionally use the 0.05 level as an arbitrary cut off: a significance level below this is generally said to be sufficiently unlikely to have happened by chance that some other explanation is required. Bear in mind that:

- If the 'sig level' is greater than .05 you can dismiss the result as being within the amount of random variation expected. Alternatively, (particularly if the 'sig level' is not much above the arbitrary .05) you can interpret it as 'suggestive'.
- If it is below .05 then you probably should interpret the value added performance of that group of students as being significantly above or below the norm. However, to what extent you as the teacher should take the credit (or blame) for it is very much open to argument, especially if it is based on results from fewer than three years with a minimum 'equivalent no of students' of 10 in each.

Robert Coe, April 1997.

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT
Durham University School of Education, Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message);
Fax: 0191 374 3506; Email: r.j.coe@durham.ac.uk

14 May 1997

Dear colleague,

A few weeks ago you should have received two sets of information, the first showing target grades for your 1997 classes, the second showing value added performance and attitudes of students you have taught in the previous three years. I would very much like to know what you thought of this feedback, and would appreciate it if you would complete the questions below and return this form to me as soon as possible.

1. Name: ……

2. How long have you spent so far reading or thinking about each part of the feedback?

	less than 5mins	5-20mins	20mins-1hr	more than 1hr
Targets 97:	❑	❑	❑	❑
Results, graphs, etc for 94-96:	❑	❑	❑	❑

3. How much more time do you expect to spend on each?

	less than 5mins	5-20mins	20mins-1hr	more than 1hr
Targets 97:	❑	❑	❑	❑
Results, graphs, etc for 94-96:	❑	❑	❑	❑

4. Have you discussed any part of it with colleagues? ❑ Yes ❑ No

5. How easy to understand have you found it?

very easy	easy	moderately hard	hard	impossible
❑	❑	❑	❑	❑

(Please make comments about any specific parts overleaf.)

6. How would you rate the usefulness of each part?

	extremely useful	useful	of some use	no use at all
Targets 97:	❑	❑	❑	❑
Student results:	❑	❑	❑	❑
Graphs:	❑	❑	❑	❑
Class averages:	❑	❑	❑	❑
Summary by teacher:	❑	❑	❑	❑

7. Please add any other comments overleaf.

Thank you very much for your help with this.

Yours faithfully

Robert Coe

Durham University/ALIS Feedback Project

Durham University School of Education,
Leazes Road, Durham DH1 1TA

Questionnaire

Summer 1997

Name: .....

Please rate the following statements on a five point scale from 'agree strongly' to 'disagree strongly':

	agree strongly				disagree strongly
A. I like to receive objective feedback about the quality of my work.	❑	❑	❑	❑	❑
B. I am always keen to have my performance assessed.	❑	❑	❑	❑	❑
C. The exam results of the students I teach reflect my ability as a teacher.	❑	❑	❑	❑	❑
D. I believe I am a good teacher.	❑	❑	❑	❑	❑
E. I do not like situations in which I am being judged.	❑	❑	❑	❑	❑
F. If ALIS gave me information about my teaching performance I would find it useful and informative.	❑	❑	❑	❑	❑
G. The ALIS data on attitudes do not tell us anything worthwhile.	❑	❑	❑	❑	❑
H. I am responsible for the exam performance of my students.	❑	❑	❑	❑	❑
I. The value-added scores (residuals) calculated by ALIS are a fair way of measuring how well students have done.	❑	❑	❑	❑	❑
J. I am worried that feedback about my teaching performance could be used against me.	❑	❑	❑	❑	❑
K. I often have doubts about whether I am doing a good job.	❑	❑	❑	❑	❑
L. The quality of my teaching is reflected in the exam success of my students.	❑	❑	❑	❑	❑
M. My institution gets very little benefit from being in ALIS.	❑	❑	❑	❑	❑
N. I am concerned that information from ALIS could be used to check up on me.	❑	❑	❑	❑	❑
O. I feel anxious when I am evaluated.	❑	❑	❑	❑	❑
P. The A level grades that students get depend on who teaches them.	❑	❑	❑	❑	❑
Q. I feel confident about the quality of my work.	❑	❑	❑	❑	❑
R. If ALIS gave me information about my teaching performance I would find it quite threatening.	❑	❑	❑	❑	❑
S. I think the Head/Principal should not use ALIS results in staff appraisal.	❑	❑	❑	❑	❑

Please describe any changes which may have resulted from your involvement in this project, specifically:
1. Changes in your attitude towards ALIS and the feedback it provides:

2. Any changes in how you will use ALIS feedback in the future:

3. Any changes in your teaching:

Would you be in favour of ALIS providing (in addition to what is currently sent) feedback on the performance of each class

	Yes	No	No opinion
sent only to the individual teacher(s) involved	❑	❑	❑
sent to the Head of Department	❑	❑	❑
sent to the ALIS coordinator in each institution	❑	❑	❑

Do you have any other ideas about how the feedback from ALIS could be improved?

Thank you very much for completing this questionnaire.
Please return it in the enclosed s.a.e.

Robert Coe

Introduction:

'This interview should take about ten minutes. I hope it will enable me to get a better understanding of your attitudes towards ALIS and feedback in general and to check whether my interpretation of what you said in the questionnaire is correct. (If I have sent you some feedback, I would also like to explore in more detail what you thought of it.)

'I would like to record our conversation if you are happy about that?' (pause) 'Everything you say is, of course, confidential.'

If received feedback:

'Which bits were easy to understand?'

'Were there any parts of it you found hard to understand? If so, why?'

Triangulation (semantic differential):

'I have some questions which I want you to try to answer using a scale from 0 to 10, so 5 is the middle value in each case.'

'To what extent do you perceive your students' success or failure as being within your control? where 0=nothing to do with me; 10=totally in my control' … 'Do you think your feeling of control has changed over the last year? If so, why? (Has the feedback I sent had any effect on it?)'

'How confident do you feel about your effectiveness as a teacher? where 0=not confident at all; 10=extremely confident' … 'Do you think your feeling of confidence has changed over the last year? If so, why? (Has the feedback I sent had any effect on it?)'

'To what extent do you believe the information provided by ALIS to be a fair measure of performance? where 0=totally unfair; 10=entirely fair' … 'Do you think your view of its fairness has changed over the last year? If so, why? (Has the feedback I sent had any effect on it?)'

'How would you describe your general attitude to ALIS? where 0=very negative; 10=very positive' … 'Do you think your attitude has changed over the last year? If so, why? (Has the feedback I sent had any effect on it?)'

Appendices for Chapter 6:

Analysis and Interpretation of findings from Project 1

(names have been changed)

Brian, Head of Maths, 11-18 Comprehensive.

I Could you tell me what kinds of feedback you think you get about your own teaching?

B Do you mean officially?

I All kinds: official or unofficial, formal or informal, immediate or long delayed. Really I'm interested in everything.

B Well we have an appraisal system and lessons are viewed and comments are made about that. Most schools have – well you have to have – appraisal, but they may change their models of it slightly. We get feedback from parents' evenings, we get parents in. If their child's enjoying it – or if they're not enjoying it – they're very quick to tell you that. Sometimes the headmaster might mention something which they've heard. But not really all that much. I think you can live in vacuum.

I So is that something you are conscious of, not actually getting much feedback?

B Yes, because once you stop being an NQT, you don't get very much feedback at all, or you could go with very little at all . I suspect the culture of that is changing with appraisal.

I What about value added feedback, is that something you get?

B Value added from a personal point of view hasn't been. We've been given departmental information and school information. But I'm actually looking after that now and I've been to heads of dept in other subjects and said what do you do with your information. Some people have said 'nothing' some have 'I don't know what to do with it' and others have said that they have actually used it quite a lot . We as a department have looked at it from a departmental point of view, but not from a personal point of view. All departments that I have interviewed are keen to get personal information.

I That's interesting, so you have actually raised that issue and people want to have that?

B Yes. Some people don't want me to provide it myself, but to provide them the data from which they could glean that information and to show them how to use Excel and things.

I On the basis of all that feedback, then, I'm interested in how you decide how well you are doing as a teacher. Do you have a feeling about the quality of your work, and is it based on any of that information you have described?

B I think as we get that information it will inevitably be part of it, because we tend to believe statistics. I suspect the people who believe it most will be those who are least happy with numbers, so we may take it a little less as gospel than some of the other departments who are less statistical. But I think a lot of realistic teachers tend to know what's going on. They do know how they're doing without getting specific feedback. But I'm sure this will have an impact.

I How do they know? Do you have feeling about that?

B Gut feeling mainly, isn't it? It's the reaction of the classes, and in parents' evenings, and of course exam results as well, but you don't necessarily have exam classes every year.

I OK. Are there any other kinds of feedback that you would like to get?

B I don't know … I'm afraid I'd have to say I don't know, not having thought about that one.

I OK. Can you describe any effects that having any particular kinds of feedback has had on your attitudes or feelings about your work?

B Well, we had an OFSTED inspection and of course you get feedback from that. We had a very good inspector who gave debriefs on every lesson immediately, which was unusual, because not everyone did that from that team. That was very helpful, very useful. We trusted and respected his opinion, and so if he said that he wasn't very happy about something then we were quite happy about that, because we believed and trusted him. On the other hand, some of the lessons I thought … I was viewed twice and one lesson was absolutely excellent and the other was … it was OK. But I found it a little boring and he was quite happy with both, so I'm not quite sure.

I That was going to be another question about the credibility you give to the various different sources of information. Supposing two sources conflicted?

B Well, that's inevitable in education.

I Absolutely. So which would you believe?

B Well, I think I know … really. Some things may be pointed out, but I still think I know, and I think a lot of teachers know what is going on in their lesson. They can tell by the feel of it. The same way that I can walk in to a lesson and I think I know straight away if it's good or bad – there's an atmosphere. That may be partly what the inspector was working on.

I If you were asked to justify that feeling, would you be able to do that?

B That's hard. You'd be able to pick out certain things. The inspectors use certain criteria, such as whether the children are all on task, and that sort of thing, and that would be part of it. I think you can tell by watching what the children talk about. I think that's an important indicator. I particularly enjoy it when I have noisy lesson where the children are arguing about their maths, that is good. Somebody walking past may think 'that's a noisy lesson' but if they came in they would soon pick up what the noise was about. So I think discussion amongst pupils about what they're doing is very important.

I So that's a kind of immediate feedback, that you're getting at the time.

Tim, Head of Maths, 13-18 Upper School

I Could you tell me what kinds of feedback you are conscious of getting about your own performance?

A Er …

I I'm interested in all kinds, a very broad range of feedback.

A Not necessarily in terms of value added data?

I Not necessarily, no, but if that is appropriate …

A In general through my senior team links – each member of staff is liked with a senior member of staff and we have senior team link meetings – and they generally comment on our views of school and what's going on in school and our performance through that way. In terms of departmentally, obviously we discuss that type of thing within departmental meetings. As a head of department it's difficult giving oneself feedback, although I get that through the deputy head who is my senior team link. In terms of the value added data, I do all the value added data for the school so I give myself feedback, as it were, in terms of how the department are performing. I recently – well a few months ago – we set targets in terms of the percentage of A*-C grades for each department, and obviously the maths department is included in that. I look at things like the value added for each teaching group and departments ask me for value added figures for their teaching groups, both at GCSE and A level. So, very briefly, that's how we monitor performance. Obviously there's a lot more we do as well.

I If different sources gave you different or conflicting information about performance, how would you rate their different credibilities?

A It doesn't often give you conflicting information. As you well know, with the value added data, it often confirms what you already know anyway. There are some discrepancies, anomalies maybe, but not many of them. There are a few, but not many. It often confirms what you already know but you've got some concrete figures to back it up. That's the beauty of the value added data: not just basing things on gut feeling.

I Do you have a perception of yourself as being a good teacher or a good head of department, and is that based on any particular feedback that you have had?

A I'd say yes, I am a good head of department and a good teacher as well. I've been appraised by the deputy head, so through the appraisal procedure and the way the department runs and the feedback I get from the members of the team – the way we're progressing in mathematics. One basic figure we look at is how many of our students go on to do A level mathematics. When I first came here just over three years ago we had maybe half a dozen in each year group doing A level, we now have twenty in each year group. There's a number of reasons for that. And it is the most popular A level subject in the school by far. Obviously it's only a crude measure, but its a measure.

I So that is a kind of performance indicator which is a kind of feedback measure, and it sounds as though you're also talking about more informal kinds of day-to-day feedback that you get about the running of the department and the way students are behaving and teachers and so on.

A That's right. We're a very informal department. We have a common room-office in the department and we meet informally here every single break-time and most lunch-times and a lot of the time is spent discussing what goes on in the classroom informally and discussing pupils, so we don't need to formally discuss 'best practice' and that type of thing within departmental meetings, as other departments do, because we get together so often informally and that's how we build up good practice and get feedback form each other about what's going on.

I One thing you haven't mentioned is any feedback you get from parents.

A Feedback from parents. The main way is through parents' evenings. Each pupil does have a personal planner which if used correctly is filled in each day and at the end of the week is checked and signed by the forma tutor, is checked and signed by the parents and there's space for the parents to comment as well. But as you well know, it's used to varying degrees and when it's used well its brilliant, and when its used badly it's not used at all. We introduced this year something called 'Discipline for Learning' into the school to try to get to grips with some of the problems we have with some of our difficult pupils, because its not the easiest school to work in. We're an inner city school in a big estate. Discipline for learning is a way of positively rewarding pupils behaviour, and obviously there are consequences for poor behaviour, but one of the positive rewards is to actually come in to contact with the parents over the phone to invite them to phone you to praise their son or daughter. That has worked quite well when teachers have remembered to send the little card to say please phone me at such and such a time, and we need to be better at that so we're going to have a re-launch in September and people are going to be encourage to contact parents through that way. Because we've found that often the only contact with parents we have – apart from parents' evenings – is when we have them in because they're excluded or have done something wrong, rather than done something good or positive.

I That sound an interesting scheme. One last question. Are there any other kinds of feedback which you don't get but would like to have about your performance?

A We certainly get a lot of feedback from the pupils, you get that whether you like it or not … Again going back to parents' evenings, often you see the parents of pupils you don't really want to see – the good ones – and the ones you'd like to see, not because you want to say how bad they are, but just to tell them how concerned you are about their mathematics, and its often those who don't come. Getting feedback from them is vitally important, particularly in a school like B. If you look at YELLIS data, we are certainly not an average school. I think in our present year ten, we have only 9% in band A [top quartile nationally] and going down to about 48% in band D [bottom quartile], so we are skewed very much toward the bottom end. That's the beauty of the value added data, because we don't expect 50% of our pupils to get 5 A*-Cs because the potential isn't there, but we do expect to get better than we do at the moment. Hopefully, OFSTED will take notice of that and not just of the crude 5 A*-C percentage.

I Have you had an OFSTED inspection?

A Just over two years ago.

I And was any of the feedback that you got from that …

A We weren't in ALIS and YELLIS then.

I They presumably gave you some information about how well they thought you were doing?

A They did, yes. We didn't fail, but in a number of crucial aspects I think we were close to failing, because of our A*-C percentage. There are five measures, one of them being the A*-C percentage, well ours is under 20% – their magic figure is 20%. The number of exclusions is another and the number of absences – unauthorised absences. Being a 13-18 school we suffer a little bit compared to 11-18 or 11-16 schools, because the older they get the more likely they are unfortunately to have absences which aren't accounted for. So in those things that are sometimes difficult to do anything about quickly you have no control and could be deemed as failing, so we were close to that last time. With the changes in the procedures, on the new criteria, if we were inspected tomorrow, we may well be close to failing again. But I certainly feel in two years since we've had the inspection, the school has gone forward a long, long way and moved forward – it is a moving school.

I Are you saying, then, that the judgement that OFSTED might make wouldn't take sufficient account of certain factors about the school?

A We would make sure now – obviously I would be responsible for presenting all the value added data – I would make sure now that all the value added data is presented in such a way to put the school in the best light, that we are moving forward, that we are making progress. Yes, if you use those crude figures, we may have not moved very far, but if you look at the value added figures, we have gone forward in a lot of key areas. It's trying to present the school in the best light.

Peter, Head of Languages, 11-18 Comprehensive

I Can you tell me in general what kinds of feedback you get about your own performance in your various roles? I'm defining feedback in a very general way, almost anything from formal to informal, immediate to very delayed …

P Formally, feedback tends to come via the appraisal process. It's more often than not the line manager, either another head of faculty like me, or more often than not, the senior tutor, as it were the next rung up, who would do appraisals. They are reviewed every year, we set targets every year and formal appraisal takes place every two years, so that's one way in which formal appraisal is given. All members of staff have an annual formal interview with the head. As a head of department who has been here more years than I care to think, you often get very frank comments from the head as well. Peers, other heads of faculty that you meet with formally or informally, more often than not informally – it might be over a drink or whatever – will chat to you about things they pick up. But more importantly than all of that, in my view, is my own department. We try to keep things as open as possible at half-termly meetings of the whole languages team. There is no criticism or critique as such, but there are certainly shared opinions on ways in which things would improve, or could improve, so I suppose that in a way is a form of feedback on my performance. To talk in terms of 'performance' sounds a bit dubious, but I think you know what I mean.

I I do. You're saying that you get information about how well you're doing from those kinds of comments that people make.

P Yes, in very general terms. Nobody will say that X in Spanish has produced some fabulous exam results, or anything as specific as that. Clearly in public examinations the figures and facts are there for all to see, and what I ought to mention is that the head does make a point of seeing in September everyone who runs a subject to go through the results of the public examinations, and will be quite frank and open and forward and will also set targets then. Targets, to an extent are very ambitious, but there again, he believes in pushing and that's fair enough. I like to push him back sometimes, but there we go.

I And things like exam results, then, in whatever form you have those, would you regard that as being some kind of feedback on your own performance?

P To an extent, yes, inevitably. If you're in charge of a big department, like languages which is big at this school, then clearly one performance indicator is the examination result. We are lucky in that so far, since I have been in charge they have been very respectable, if not very good on occasions, so the feedback has been positive. I have yet to be confronted with negative feedback about exam results, but no doubt one day that will come. We all have off years, as it were.

I What about any other sources of feedback? You haven't said anything about students, parents …

P I was just going to come on to that. The student feedback is not direct, but you can pick things up about me, or about other members of the dept, which you could if you chose to as head of dept, turn a blind eye to. But if I hear something going on that's serious, then in as subtle a way as possible I'll try to address that. And parentally , the best time for feedback is the parents evening, but they tend on the whole to be very positive. The sort of parents that you'd like to have in because their child is presenting whatever difficulties for the department tend not to come to parents evenings. Negative feedback is not impossible or a non-event in parents evenings, but fairly infrequent.

I So what other kinds of feedback would you like to get that you don't? You've mentioned one there.

P I would like more formal feedback from the parents. At the end of the day, the school will stand or fall based on what parents in the local area think. Clearly the grapevine has a fairly strong influence on who sends their children to the school. Because we are in a small city with lots of good big comprehensives like this, then whether we like this or not, and sometimes we don't, we are in a competitive area and we have to therefore please the people we are aiming at. So certainly, more positive feedback from parents. Other than that, I can't actually put my finger on one particular area that I'd like more feedback from, because in terms of what goes on in school there is a fair amount that comes our way.

I OK. Are there differences between those sources in terms of the credibility you give them? Possibly if two sources conflicted, and gave you information …

P I know what you're getting at. If we were to do a direct comparison and say for the sake of argument a criticism of a member of my department from a parent, whether it was to me in writing or verbally. Or, put it the other way round, if I'd heard some tremendous praise for a member of my department from a parent, and yet the same member of my department was criticised, and I heard about it from students, I might not give the student voice there a lot of time. I'd probably want to just check it out. I think whether it's right or wrong, you do tend to regard parental feedback, whether it's criticism or praise, more positively – more seriously – than you would if it comes from students, although that's not to say that I'm happy to dismiss what students think, because clearly they are as important in this as anyone else. Obviously, if the head stops me or calls me into his room and wants to tell me that somebody is doing extremely well or isn't doing extremely well then that would probably carry most weight.

I How important is feedback to you personally?

P About me or about my department?

I Whatever you would term your own performance, to the extent that you feel responsible for the department?

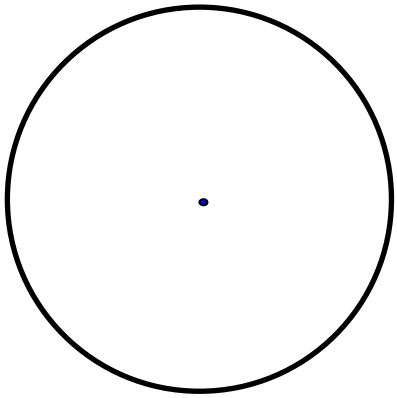
P I think it's very important. Since I've been head of languages I've been appraised twice and I've found both appraisals extremely useful, from the point of view of praising what I do, but also criticising some of the things I don't do, or rather criticising me for not doing certain things. You learn from that and I've taken a lot of things on board since then, On the whole, the things I've done as a consequence of appraisal I absolutely agree with, so there has not been any conflict there with the senior tutor who has appraised me. Far from it. We get on very well, and that's one of the reasons he's my appraiser.

I So there's a mixture of positive and critical there? How does that affect your feelings about what you do?

P I'd rather be told, to be honest. I'd rather not have something skirted around, if there's something that it is perceived that I'm not doing awfully well, I'd rather be told straight, so that I can address the problem directly, rather than someone suggest that maybe … I think a lot of the people in this school for whom I have respect, if you ask them, they will be very straight with you. It makes it sound like this is happening all the time; it isn't, but …

(End of the tape. Last few seconds of the conversation not recorded.)

Question	Variable	Coding of Responses
A2: Sex	SEX	1 = M, 2 = F
A3: School/College	INST	1 = Inst1, etc (see Ch??, 'Methodology' for description of each institution)
A4: Position	POSN	1 = subject teacher/lecturer 2 = Dep HoD/subject responsibility 3 = Head of Department 4 = Head (or Dep Head) of Faculty 5 = Senior management 6 = other *(Highest coding used if ambiguous)*
A5: How long have you worked there?	YRS	Time in years (to nearest 0.5)
A6: Which of the following is your main subject taught at A level?	SUBJ	1 = English 2 = French 3 = maths 4 = physics
A7: Do you teach any classes which will take A level in this subject this year (97)?	EX97	1 = yes 2 = no
A8: Did you teach any classes last year which took A level in this subject in 96?	EX96	1 = yes 2 = no
B1 to B29 (Likert scale items)	B01 to B29	1 = agree strongly 5 = disagree strongly *(2, 3, 4 for in-between values)*
C1: Approximately when did you first become aware of the existence of ALIS?	AWARE	1 = within the last year 2 = between one and three years ago 3 = more than three years ago
C5: Please tick any of the following which describe(s) the stage you are at with using ALIS:	STAGE	1 = I have not had any contact with it 2 = I have begun to learn about it 3 = I have made some plans to use it 4 = I have used it in ways that have been required of me 5 = I have made my own routine use of it 6 = I have applied it effectively to solve problems 7 = I have integrated its use into my work 8 = I have evaluated and modified it to meet my needs *(Coded as highest ticked)*

D: Imagine that the circle represents the exam performance of typical students. Please divide it into sections (like a pie chart) where the size of each piece indicates the **relative importance** of that factor in determining exam performance. Factors which affect exam performance: A: students' ability B: students' home background C: students' character attributes (eg, hard working/lazy) D: which teacher they have E: which school/college they go to F: other factors (if any particular ones, please list them)	 RESP_ABL RESP_BGD RESP_CHR RESP_TCH RESP_SCH RESP_OTH	*(The size of each portion of the circle was determined by the arc length at the circumference. This was coded as the percentage of the total. Measurements were checked if total percentage was not within ±2 of 100%.)* 

SEX	Frequency	%	Valid %
male	34	47.2	47.9
female	37	51.4	52.1
missing	1	1.4	

INST	Frequency	%	Valid %
Inst1	13	18.1	18.1
Inst2	2	2.8	2.8
Inst3	17	23.6	23.6
Inst4	15	20.8	20.8
Inst5	3	4.2	4.2
Inst6	5	6.9	6.9
Inst7	6	8.3	8.3
Inst8	6	8.3	8.3
Inst9	5	6.9	6.9
missing	0	0	

POSN	Frequency	%	Valid %
subject teacher	31	43.1	44.3
Dep HoD/subj responsibility	7	9.7	10.0
Head of Dept	17	23.6	24.3
Head of Faculty (or Deputy)	8	11.1	11.4
Senior management	2	2.8	2.9
other	5	6.9	7.1
missing	2	2.8	

SUBJ	Frequency	%	Valid %
English	21	29.2	30.0
French	7	9.7	10.0
maths	29	40.3	41.4
physics	13	18.1	18.6
missing	2	2.8	

EX97	Frequency	%	Valid %
exam class for 97	66	91.7	94.3
no exam class	4	5.6	5.7
missing	2	2.8	

EX96	Frequency	%	Valid %
exam class for 96	61	84.7	87.1
no exam class	9	12.5	12.9
missing	2	2.8	

AWARE	Frequency	%	Valid %
within the last year	2	2.8	2.8
between 1 and 3 years ago	25	34.7	34.7
more than 3 years ago	45	62.5	62.5
missing	0	0	

STAGE	Frequency	%	Valid %
1 (no contact)	2	2.8	3.0
2 (begun to learn)	12	16.7	18.2
3 (made some plans)	3	4.2	4.5

4 (used as required)	23	31.9	34.8
5 (own routine use)	13	18.1	19.7
6 (applied to solve problems)	0	0.0	0.0
7 (integrated into work)	4	5.6	6.1
8 (evaluated and modified)	9	12.5	13.6
missing	6	8.3	

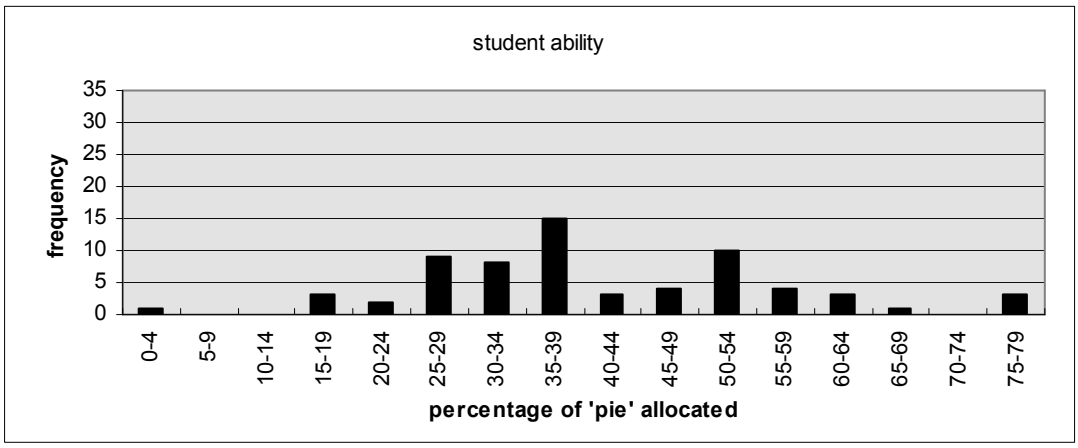
Statement	agree strongly				disagree strongly	
coding:	**1**	**2**	**3**	**4**	**5**	**missing**
B1. I like to receive objective feedback about the quality of my work.	32	29	9	2	0	0
B2. I am always keen to have my performance assessed.	8	29	27	7	1	0
B3. The exam results of the students I teach reflect my ability as a teacher.	6	23	33	7	3	0
B4. I believe I am a good teacher.	15	45	10	0	0	2
B5. I do not like situations in which I am being judged.	4	15	28	19	5	1
B6. If ALIS gave me information about my teaching performance I would find it useful and informative.	18	25	22	4	3	0
B7. My effectiveness as a teacher depends on how I choose to teach.	13	39	15	3	0	2
B8. Receiving feedback can help me to improve what I am doing.	14	50	7	0	0	1
B9. If a student who does not much like my subject joins my class, I can usually help him or her to enjoy it more.	4	29	37	2	0	0
B10. The ALIS data on attitudes do not tell us anything worthwhile.	7	10	33	15	3	4
B11. I prefer tasks in which I can see how well I am doing.	4	35	29	3	1	0
B12. I am responsible for the exam performance of my students.	7	24	29	11	1	0
B13. The value-added scores (residuals) calculated by ALIS are a fair way of measuring how well students have done.	5	23	28	14	2	0
B14. I am worried that feedback about my teaching performance could be used against me.	3	18	26	20	5	0
B15.I often have doubts about whether I am doing a good job.	5	12	20	28	6	1
B16. If the students I teach perform badly, it is their fault.	1	10	39	20	0	2
B17. The quality of my teaching is reflected in the exam success of my students.	3	33	27	7	2	0
B18. I usually seek information with which to judge whether I am achieving what I want to.	3	26	29	12	1	1
B19. My institution gets very little benefit from being in ALIS.	3	10	34	16	6	3
B20. I am concerned that information from ALIS could be used to check up on me.	3	15	22	22	10	0
B21. I feel anxious when I am evaluated.	5	23	25	18	1	0
B22. The A level grades that students get depend on who teaches them.	3	24	28	14	3	0
B23. There are too many errors in the feedback provided by ALIS for their findings to be reliable.	1	19	35	11	2	4
B24. If the analysis by ALIS shows that a particular department has a high score, then I will believe that there has been some good teaching in that department.	8	33	25	6	0	0
B25. I feel confident about the quality of my work.	12	47	10	2	0	1
B26. If ALIS gave me information about my teaching performance I would find it quite threatening.	2	6	26	34	4	0
B27. Doing well is more important to me when I am being assessed.	2	12	24	22	10	2

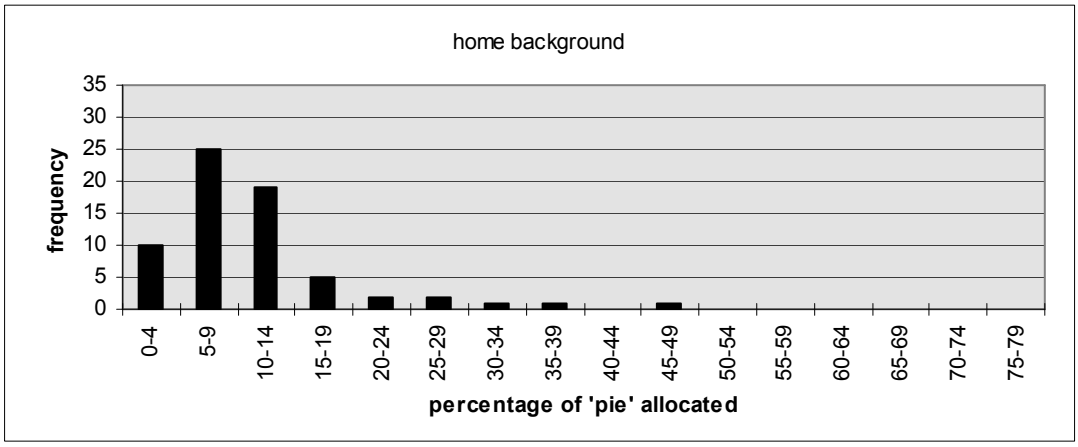
B28. When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough.	1	26	12	16	17	0
B29. I think the Head/Principal should not use ALIS results in staff appraisal.	15	17	21	14	2	3

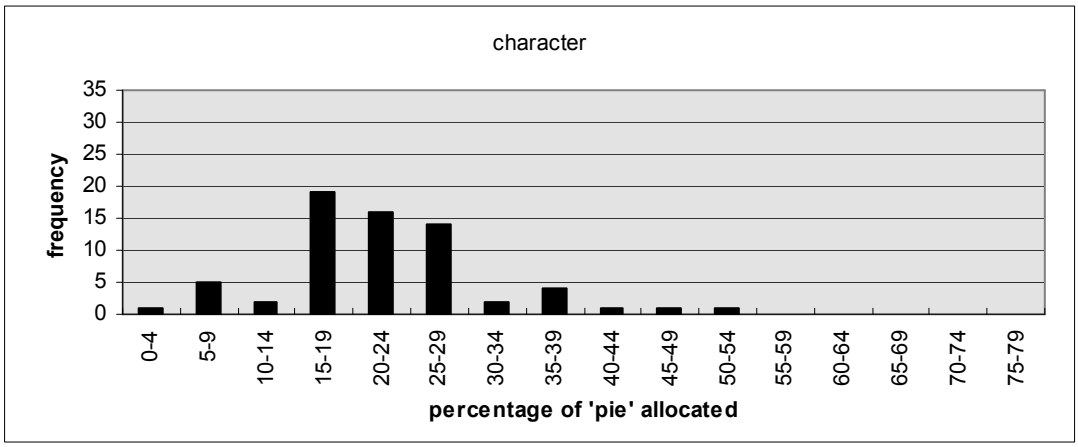
Statistics for each of the 'pie chart' factors:

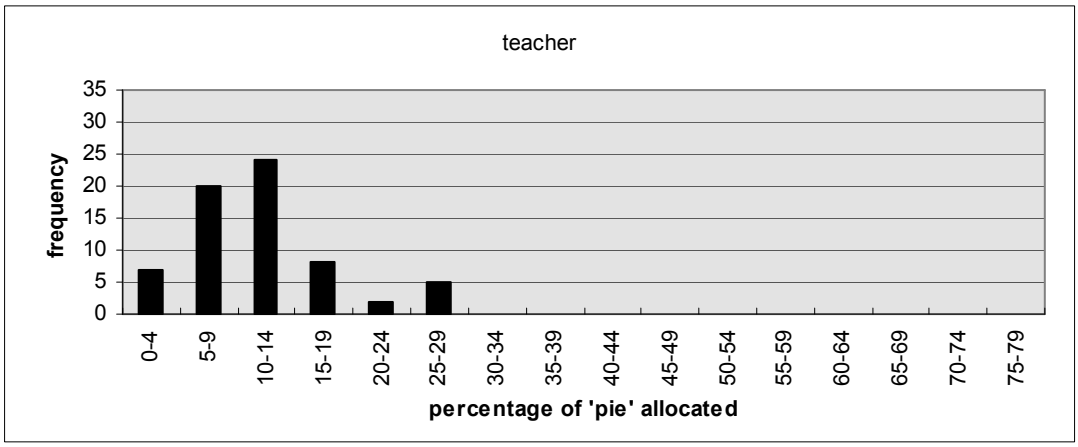
	student ability	home background	character	teacher	school	other
n	66	66	66	66	66	66
min %	0	0	0	1	0	0
max %	75	48	50	28	23	22
mean %	40.5	11.2	21.7	11.5	8.4	6.0
median %	37.5	9	21	10.5	7	5

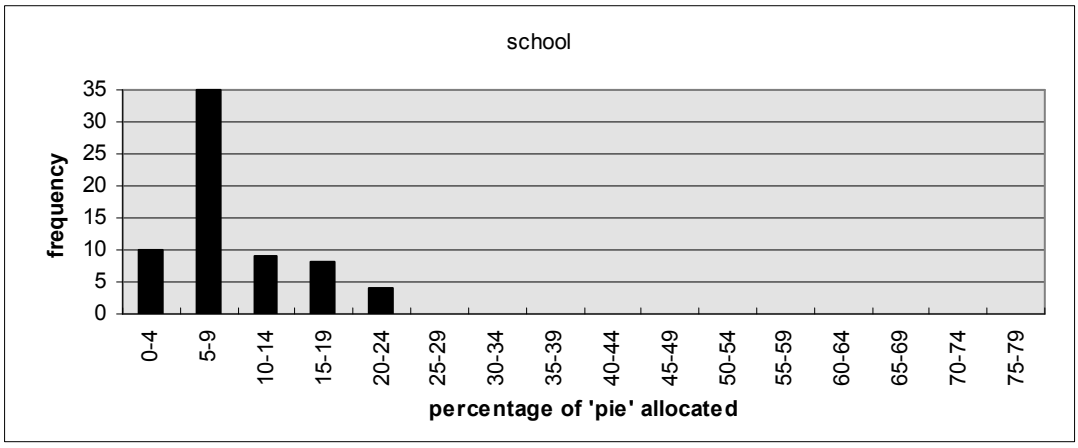
Bar charts showing the distribution of percentage of 'pie' allocated to each factor:

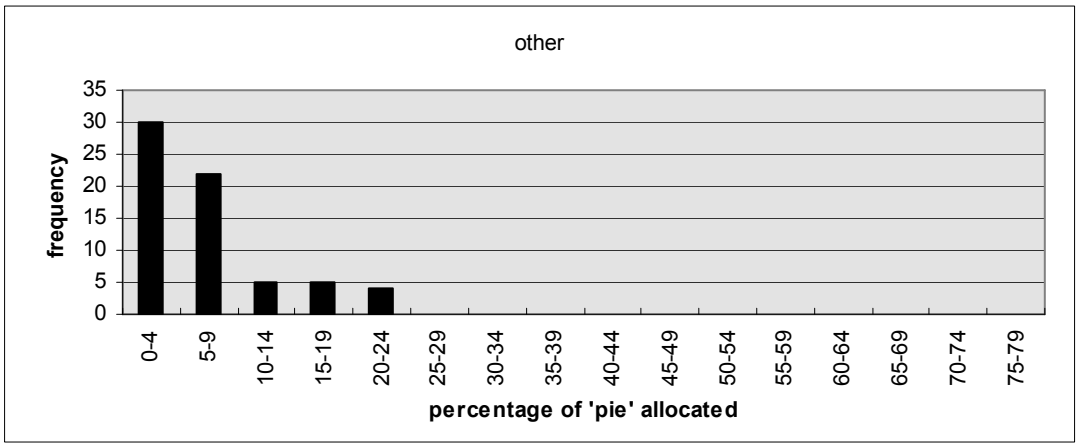





character
frequency
0
5
10
15
20
25
30
35
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
percentage of 'pie' allocated


teacher
frequency
0
5
10
15
20
25
30
35
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
percentage of 'pie' allocated


school
frequency
0
5
10
15
20
25
30
35
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
percentage of 'pie' allocated


other
frequency
35
30
25
20
15
10
5
0
0-4
5-9
10-14
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
percentage of 'pie' allocated

B: Please list any forms of feedback (formal or informal) or information you have had about your performance in this job:

2/1 Appraisal, EQR, Feedback from Vice-Principal
2/2 Collation of standardised residuals for individual students, and hence group averages etc
3/3 None as yet
¾ Appraisal, students comments, informal student questionnaires
3/5 So far - v. Little. Discussions with HOD and other staff (always informal) have allowed me to harmonise my teaching/marking/planning to the house style which was useful. No formal information received as yet
3/7 I had feedback when I was on probation, during inspection, ALIS and from some review sheets the students have to complete during their course
3/8 ALIS
3/9 ALIS results and A level results
1/10 ALIS analysis
1/11 Discussions with line-manager/appraiser, though these have concentrated on aspects other than teaching. Feedback from inspections - HMI, [county] and EQR - these have mostly been departmental, though some personal comments have reached me through the principal. Indications from student performance.
3/12 PRAD, response to Departmental Review.
1/13 Staff review (appraisal), Principal interview, student evaluations.
3/14 Appraisal by HOD and Senior Tutor.
3/15 Formal: Inspection (dept awarded grade 1), appraisal; Informal: ITT students observing lessons.
3/16 Appraisal.
4/17 Induction year report from Head of Dept.
4/18 Discussion of methods used.
4/19 Informal chat with Principal re results, appraisal (theoretically).
3/20 Formal staff appraisal took place but didn't really give me any information or feedback. We ask the students to fill out questionnaires about their attitudes to the course/teaching etc. & discuss results of these. We've looked at ALIS feedback too.
*5/21 Appraisal, student perception of course, ALIS, retention rates, exam pass rate.
4/22 An appraisal in 95.
3/23 Routine course reviews completed by students. Indirect comments via Records of Achievement. Professional Review Development.
4/24 Very little - except that some texts taught have been more popular than others.
3/25 Student feedback - oral & from course review.
4/26 Informal via other teachers from student comments.
6/28 Virtually none from the [combined sixth form centre] management. My own head at [school] has commented on the success of exam candidates this year. ALIS was also used to point out that some [subject] students had stated they disliked [subject] on their ALIS return. This was pointed out at a large management/faculty meeting, negatively, and in my view inappropriately.
6/29 Informal discussions with colleagues/more senior staff, exam results, examiners' reports.
1/31 Appraisal, inspection from outside inspectors & Head of Department sitting in on lessons.
4/32 Appraisal last June.
5/33 Student Perception of Course returns.
4/34 Appraisal (but yet to be done!) Discussion with Principal and Director of Curriculum re previous year's results.
4/35 Full inspection, appraisal
6/36 Formal and informal discussions with head of English. OFSTED inspection lesson observation. Informal discussions with other departmental members.
4/37 Eng. Dept. meetings to go through ALIS findings.
5/38 Staff appraisal (formal). Open testimonial written in connection with job application (Senior Tutor). Informal comments.
1/39 No formal feedback apart from one GCSE oral assessor who was complimentary about a lesson she observed. Informal feedback = occasional thanks / compliments from A level students at end of course.

1/41 Feedback from C[urriculum] M[anager].
1/42 Student evaluations
1/44 review discussion with appraiser (line manager). External quality review (although feedback was given in general terms). Student evaluations.
3/45 Can't think of any.
6/46 ('your performance' underlined) None - not directed at me personally.
4/47 Every year as a department we calculate the average residual for each teaching set.
7/48 FEFC inspection (Oct 96) (Subject area achieved grade 2)
3/49 Staff appraisal. Parents thanking for the good exam results. ITT students observing lessons. FEFC inspection.
9/51 The only feedback I have received is through the A level results. All students passed, estimated grades were accurate or improved on.
9/52 Appraisal. HMI inspection (on department's performance). Informal comments from Head, colleagues, pupils, parents.
7/53 H of D (in 1st yr). Students and student questionnaire.
1/54 Informal students' comments. EQR inspection by members of staff from another college.
8/55 Appraisal. Thanks from Head of Dept / parents / pupils.
8/56 Appraisal. Chatting with fellow teachers
8/57 Appraisal - formal. From Head of Department - informal.
9/59 Appraisal. Inspection. Comments from: NQTs and student teachers, staff and deputies, students and pupils I teach. Exam performance and ALIS?
8/60 Positive feedback about performance with groups from years 7 to 9 from department head, department colleagues and parents. No feedback at all for years 10 to 13.
4/61 1. from the Principal once a year in discussion of exam results; 2. from the ML inspector at the time of the college FEFC inspection; 3. from on of the senior managers as part of my appraisal.
8/63 Pupil comments of satisfaction and appreciation!
4/64 Appraisal. ALIS data. Parents' evenings. Students' comments.
3/65 From whom? The formal appraisal (done by VP) was very positive. Some students are grateful but obviously, most aren't or are too 'cool' to express it if they are.
1/66 Annual review by curriculum manager. Student evaluations.
9/67 As a department, how our results match the ALIS predictions.
7/69 Informal. Staff meetings. Inspection report 1996.
9/71 Formal review of department progress by curriculum sub-committee - positive comments. Pre-OFSTED inspection - detailed and again positive feedback. Otherwise, in school, NONE!

B (Comments added to Likert scale items):

B1. I like to receive objective feedback about the quality of my work.

4/50 ('objective' circled; coded 4)
6/28 (see comment in E; coded 3)

B2. I am always keen to have my performance assessed.

B3. The exam results of the students I teach reflect my ability as a teacher.

6/27 Not necessarily (coded 4)

B4. I believe I am a good teacher.

1/11 [main subject] only! (coded 2)
3/23 Depends on your definition of 'good' (left blank)
4/47 modesty forbids (left blank)
4/61 (see comment in E; coded 2)

B5. I do not like situations in which I am being judged.

4/26 ('judged' circled with a question mark; coded 2)

1/54 (coded 4) ie I quite like them!

B6. If ALIS gave me information about my teaching performance I would find it useful and informative.

2/1 If this were feasible (coded 1)
3/65 I don't believe it can. There are too many variables (coded 4)
9/71 If! (coded 2)

B7. My effectiveness as a teacher depends on how I choose to teach.

5/33 If 'how I choose to teach' includes homework assignments etc., I would agree more strongly (coded 3)
1/54 ('choose' underlined) within constraints over which I have no control (coded 3)

B8. Receiving feedback can help me to improve what I am doing.

B9. If a student who does not much like my subject joins my class, I can usually help him or her to enjoy it more.

5/33 Seems rather unrealistic - students who don't like a subject are unlikely to do it at A level. (Coded 3)

B10. The ALIS data on attitudes do not tell us anything worthwhile.

4/24 ('attitudes' underlined with a question mark; left blank)
8/60 NK (left blank)

B11. I prefer tasks in which I can see how well I am doing.

4/18 A whole lesson? Tasks in life in general? Poor question ('tasks' circled, coded 3)
9/71 ? (coded 3)

B12. I am responsible for the exam performance of my students.

6/27 To some extent (coded 3)
1/42 partly (inserted between 'am' and 'responsible'; coded 2)
6/46 It is a shared responsibility (coded 3)

B13. The value-added scores (residuals) calculated by ALIS are a fair way of measuring how well students have done.

1/30 (Comment written in section E:) The value-added scores may be a fair way of measuring how well the total cohort has done but not for measuring individuals. Eg, ALIS target 5.6, Achieved D, but negative residual - yet achievement was on target.
1/54 Not being a statistician, I'm not sure (coded 3)

B14. I am worried that feedback about my teaching performance could be used against me.

B15. I often have doubts about whether I am doing a good job.

B16. If the students I teach perform badly, it is their fault.

6/46 - as question 12 (coded 3)
1/54 could be (coded 3)

B17. The quality of my teaching is reflected in the exam success of my students.

B18. I usually seek information with which to judge whether I am achieving what I want to.

4/61 (see comment in E; coded 3)

3/65 You don't need to seek it! It's there in the students' work and behaviour (coded 4)

B19. My institution gets very little benefit from being in ALIS.

3/7 I don't know. (left blank). (Also added in E:) I cannot answer this question
4/24 (question mark added; left blank. See also comment at end of section)
6/46 We need to use it more widely - and have time to analyse it in detail (coded 3)
1/54 I wonder (coded 3)
8/60 NK (left blank)

B20. I am concerned that information from ALIS could be used to check up on me.

B21. I feel anxious when I am evaluated.

B22. The A level grades that students get depend on who teaches them.

B23. There are too many errors in the feedback provided by ALIS for their findings to be reliable.

1/54 Mistakes this year; any in previous years? (coded 2)
8/60 NK (left blank)
8/62 (see comment in E; left blank)

B24. If the analysis by ALIS shows that a particular department has a high score, then I will believe that there has been some good teaching in that department.

1/54 though I'm prepared to believe this is not so (coded 1)

B25. I feel confident about the quality of my work.

B26. If ALIS gave me information about my teaching performance I would find it quite threatening.

B27. Doing well is more important to me when I am being assessed.

4/18 than what (inserted after 'me') Meaningless question (left blank)
3/65 No! What's important is the quality of education we can give to the students which has deteriorated as a result of diminished resources, too many students, too much contact time and too many <u>other</u> things like laborious assessment procedures and filling in questionnaires!

B28. When I think about the weakest areas of my work, I usually feel they are a result of my not trying hard enough.

B29. I think the Head/Principal should not use ALIS results in staff appraisal.

4/24 (question mark added; left blank)
6/46 Only if initiated by appraisee (left blank)
8/62 (see comment in E; left blank)

Comments added at end of section B:

4/24 I do not know how ALIS is used in feedback

C1: (Comments added)

3/7 Maybe more than 3, I don't remember (coded 2)
1/40 I think! (coded 3)

C2: What information have you had from ALIS about the performance or attitudes of your students?

2/1 Performance info is v. thorough - though ALIS does not take into account A level syllabuses in [subject], which most A level teachers believe, influences exam results. Attitudes info is less 'user friendly' and seems much less useful - vague.
2/2 Residuals for [subject] results.
3/3 None yet
¾ Some but usually appears too general to be of use an individual teacher.
3/5 As yet none apart from general information circulated by HOD about the strength of this years' cohort when compared with last years'.
4/6 Poss to draw regression lines to see if students have performed better/worse than expected.
3/7 The value added is above average and the attitudes not bad. The students usually like [subject] when they get their good results in August!! It is perceived as the hardest A level.
3/8 Average GCSE score, residuals. Not particularly interested in attitudes.
3/9 Booklet on attitudes and residuals.
1/10 Annual Report.
1/11 ALIS tables for residuals and attitudes.
3/12 Annual reports - useful when interpreted for us by [name], a statistician.
1/13 All ALIS reports are available to me.
3/14 The students doing [subject] have consistently achieved 1 - 1.5 grades better than predicted on entry. That they don't regard [subject] as their main subject, but do it as their 2nd or 3rd A level..
3/15 Individual residuals, regression lines, 3 year moving averages, etc., etc. Attitudes to subject/college.
3/16 Residuals info.
4/17 Yearly report (attitude to subject, college, course, etc.).
4/18 Feedback on their questionnaires.
4/19 Residuals in performance (after results), control charts, calculated statistics on residuals, summary of attitudes data.
3/20 Looked at comparison between final grades and expected ones. Information has been given us by head of dept.
*5/21 Residuals for past 4 years.
4/22 Regression lines per subject, residuals for each student.
3/23 Regular information is given in departmental meetings.
4/24 Only what kind of GCSE results they had.
3/25 Attitudes to subject and to college, individual comments which refer to the subject.
4/26 None yet, as I've just joined the institution. Previous place had ALIS too.
6/27 Each group of students/student is unique: I have sincere doubts about all these comparisons - and so far have not really been convinced by the arguments for all these statistics. They have not told me anything I did not know already.
6/28 Nothing that I didn't already know although a comment made about attitudes from ALIS was not reflected in the % outcomes.
6/29 Relatively little. I have been aware of the system and how it operates but little information has come my way. I saw a report on students ~ 4 years ago but have not since.
1/30 Booklet.
1/31 Lists with results and residuals. Also for students receiving support we have looked at their residuals.
4/32 Residual values.
5/33 I assume that I've had all that the college has received.
4/34 Only the usual info/stats that ALIS produces, including isolated statements reproduced from their questionnaires.
4/35 Booklets, etc.
6/36 I have had no detailed feedback. I have begun to learn about ALIS as a member of the sen mgmt team in relation to 'promoting' the [institution] to year 11 students.
4/37 Very little that I do not already know.

5/38 Residuals for all students. Chances graphs. Averages related to ALIS score and IQ.
1/39 A few basic figures from last year.
1/40 Very little on attitude. Statistics on performance.
1/41 None
1/42 None directly (some indirectly, through curriculum manager).
1/43 Student annual evaln. of courses
1/44 All information as provided in subject books.
3/45 'ALIS' scores after final results.
6/46 Subject breakdown of responses.
4/47 Every year we receive information about performance and identify over- and under-achievers in each set. Not much info about attitudes.
7/48 The full ALIS statistics are available to all staff. As head of the subject are, I pay close attention to them.
3/49 -
4/50 Not much about the attitudes of my own students.
9/51 It can highlight under-achievers.
9/52 Performance tables. Nothing about individuals' attitudes or in any detail.
7/53 The general stuff relating to both maths and comp and faculty attitudes.
1/54 Attitude tables - but I have not really bothered with these, as I feel they are less simple to interpret than the PLRs, etc.
8/55 Only found out general comments, not subject specific ones.
8/56 None that I remember.
8/57 Residuals. Graphs of how students have done in comparison with how they were expected to do. Information on different performance between the sexes.
8/58 The Head of Department has received data from Senior Management. This has occasionally been discussed at departmental meetings.
9/59 The information about my subject is given to me.
8/60 As a teacher governor, I was aware of the existence of ALIS, but I have never actually seen the ALIS data for my department, in spite of a direct request to my Head of Department. As a result, I have obviously made no use of it and have been unable to act upon it!
4/61 1992-6: Pupil Level Residuals. 1996: Institution summary report. 1992-6: scores for attitudes to subject and college; 1995: perceived learning activity chart; 1992-5: departmental statistical control charts; 1992-5: feedback booklets showing performance and attitudes of [instn] students relative to other institutions in the cohort.
8/62 None.
8/63 Lots. They like being explained to well and being helped through old examination papers.
4/64 English is quite a popular subject. Not enough time given for students to help each other in lessons. Students likely to get grade C and below are not as well catered for as others.
3/65 Some - but it was full of errors anyway.
1/66 Scores on entry.
9/67 As a dept. - how our students should perform
7/68 (see other comments, below)
7/69 ALIS report/statistics.
7/70 Very little that I can recall.
9/71 Annual subject booklets
7/72 Value added results. Results of attitude survey.

C3: What use have you made of it?

2/1 Close analysis of ALIS exam performance a part of our examination of results - together with breakdown of grades/papers, comparisons with mock exam results and predictions and with performance of our students in other subjects.
2/2 Attitude survey (at [previous institution]) which helped to provide evidence for pastoral provision and its quality
¾ Very little.
3/5 None - I've gone ahead and taught in my usual manner.
4/6 Tried to see why some students have done badly.

3/7 None.
3/8 Little.
3/9 Usually residuals confirm view or staff and students. Useful for staff appraisal.
1/10 Very little. The information arrives too late to be of use.
1/11 In the early years I used to use the printout of student residuals to work out group averages, etc. for my results analysis. In recent years, I have needed to hand my curriculum review to [name] before the ALIS data has arrived in college. I have used the ALIS methodology to plot graphs of achieved grade against average GCSE.
3/12 We've compared predicted grades from ALIS and mocks. We've looked at how different sets performed in comparison with one another and assessed the Depts performance overall.
1/13 Much use of performance stats, both for my subject and as manager of others. Little use made of attitudes stats.
3/14 Informally tried to predict grades expected for new students from their GCSE score and use these as minimum targets. Tried to identify underachievers from past years and reasons for this.
3/15 Analyse past performance - class by class. Look for trends eg different performance by teacher, sex, etc. Confirm decision to drop a syllabus and change to another in non-cwk A level. Confirm success of [subject syllabus].
3/16 None.
4/17 Noted any change from previous years.
4/18 Discussed it, but often the findings were inconclusive.
4/19 Analysed and written a report. Fed conclusions into departmental development plan.
3/20 Not much, yet. Just found it interesting.
*5/21 Long term, little. ALIS residuals have fluctuated widely and cannot be attributed to any specific cause.
4/22 Not a great deal. Mainly to offer encouragement to students with low ALIS scores.
3/23 Mainly used and analysed by head of department who the highlights important current facets.
4/24 None.
3/25 Looked at students who did not do well - across other subjects and in relation to others with same GCSE score. Looked at residuals in groups I have taught.
4/26 (see comment in C2)
6/27 Simply noted the findings, but not over confident of them.
6/28 Very little.
6/29 None.
1/30 None - statements are contradictory.
1/31 If 'supported' students are performing better than ALIS predicted then we feel the support has been beneficial.
4/32 Analysed it.
5/33 I have used it to see whether my students are performing as might be expected of them.
4/34 None at present as the info is somewhat bland and shows nothing untoward. It might be informative on the questionnaire to ask them to encapsulate in one sentence their feelings about each subject individually.
4/35 Annual analysis.
6/36 As above.
4/37 Very little.
5/38 Chances graphs to show first years what happened previously. Residuals to display individual student scores and regression line as a single chart.
1/39 None so far.
1/40 None
1/41 N/A
1/42 N/A
1/43 General overview of performance
1/44 Have worked out average residuals for my classes and for subgroups, eg males/females
3/45 None - they are generally what we expected.
6/46 To compare with estimated/predicted grades of past students. To compare responses from year to year and subject to subject to look at reasons for justifying the outcomes. We really want something we can use BEFORE exams not after.

4/47 None
7/48 Assessing the overall achievement of a subject area - in general, the value added statistics tend to confirm our impressions of whether it has been a satisfactory or disappointing year. In general, the attitudes survey tells us little we don't already know. Occasionally, at entry we use the grade 'projections' to convince prospective students their aims are unrealistic - or try to !
3/49 -
4/50 Not much - in actual teaching - I depend upon my own educational research.
9/51 Targeting under-achieving students. Looking at performance of all students - achieving maximum potential.
9/52 Tried to compare performance tables with own expectations, reasons for these.
7/53 Checking on students (and my own) achievements
1/54 (See question 2)
8/55 Some comments have been interesting, but …. (continued in C4)
8/56 Not much
8/57 Looked at why some students have under-achieved.
8/58 Little. I find it all rather frustrating - hence my late response.
9/59 Not a lot. Lack of time, unsure how valuable ALIS info is. Not sure how to implement it.
8/60 (Blank)
4/61 Each year we have discussed the data at departmental and faculty meetings. Each year we have then evaluated the data and presented a written analysis to the curriculum director. On the basis of our evaluation we have modified or adapted programmes or methodologies as appropriate.
8/62 None
8/63 Little. It only tells you things you expected.
4/64 Tried to vary teaching and learning styles. Tried to be sensitive to the needs of so -called 'lower achievers'.
3/65 None
1/66 None
9/67 Tried to use it to identify students who are under performing.
7/68 (See other comments below)
7/69 Used for discussion
7/70 None. I am far too busy with all the admin work required.
9/71 Departmental discussion:1 Review of individual student performance 2 Consideration of teaching styles and approaches
7/72 I have used the 'grade prediction' tables to help students consider A Level study

C4: How valuable have you found it?

2/1 Valuable as part of overall analysis.
2/2 Very
¾ Has boosted morale when given evidence that students are doing better than expected.
3/5 It could throw up the existence of vastly differing abilities in my new [subject] A level classes - but so did setting a short piece of work.
4/6 Quite useful although in some cases reasons for failure lie with attitudes of student rather than ALIS score. [Subject] results at A level look more dependent on GCSE [subject] than ALIS.
3/7 Not very.
3/8 Fairly valuable to see how the students are doing as compared with other departments in other institutions.
3/9 Residuals are extremely interesting and allow confirmation or otherwise of good practice. Attitudes are not useful.
1/10 Not very. It just gives an overall view of a student's performance after GCSE but I believe that ALIS score from GCSE are not always good indicators for individual A levels.
1/11 As above. ALIS approach very useful, but actual data has not arrived until after I have needed it.

3/12 Very limited: the exceptions are usually explicable, the rest confirm what we feel to be going on. It's a serious shortcoming that ALIS does not discriminate between boards/syllabuses.
1/13 Performance stats provide a useful comparison, particularly now it is available over a 5 year period in some subjects.
3/14 Interesting and encouraging, but we have not yet used it to the full. We have plans to use it more for current students, to identify underachievers there.
3/15 Valuable first or second year - then little new after. It would be more valuable if used with current students working on Target Minimum grades and action planning.
4/17 Only interesting rather than particularly useful.
4/18 Not valuable. It is just the average of their GCSE subjects and doesn't really have a bearing on best GCSE chosen subjects, and some subjects, especially languages, are far too easy at GCSE. So most of our students have A or A*.
4/19 Quite. A useful mechanism to analyse results: but not indispensable.
3/20 Interesting but I haven't used it particularly. We haven't had anything that clearly needed acting upon.
*5/21 see above
4/22 Not particularly.
3/23 It is used along with other indicators to help assess the performance of the department.
4/24 So far, not valuable.
3/25 Reasonably.
4/26 (see comment for C2)
6/27 Usually can find faults with the statistics; eg having to demand that Maths and Further Maths are treated as separate subjects.
6/28 Not particularly. It always appears to be more relevant to those in management who like making comparisons between subjects. It gives them an overview. It has been my experience that the so called negatives it throws up are almost always more important than the positives. It does depend how it's used. My experience with YELLIS is much more positive, because of management but also because of more control over the students. Comments on attitudes are too vague and shouldn't be used to judge a subject: 'it's too difficult' etc.
6/29 n/a. I feel it could be of use to me.
1/30 Little value, some interests.
1/31 For 'supported' students very. Less so for A level [subject] students.
4/32 Fairly.
5/33 Fairly reliable. With my subject [named] there has been a slight problem in that standards have varied from board to board and it is difficult to tell the extent to which this accounts for discrepancies between the national and local picture.
4/34 A this stage it provides reassurance rather than suggests where the faults, if any, lie.
4/35 In some cases. More valuable than raw results anyway.
6/36 Of limited value - although SMT members more closely involved with year 11 to FE have found it more useful
4/37 ALIS is of interest but it cannot account for the human factor e. g. the student who stops working in u6 year, the student who has over achieved at GCSE and finds A level surprisingly hard as a result. It tells me very little I do not already know about my students
5/38 In maths, difficult to use with any conviction as there are so many exceptional and GCSE is such a poor indicator of A level mathematical performance
1/39 Not at all - so far
1/40 Not telling me anything I didn't know. I feel strongly that a regression line of confidence intervals would be much more sensible. I would also like to know the mean residual for each subject (have you standardised it to zero - only standardising the standard deviation is shown as far as I can see!)
1/41 N/A
1/42 N/A
1/43 Not as useful as internal dept evaluation
1/44 Interesting rather than valuable. Provides a more helpful indicator of student performance than raw results. Identifies those students whose performance is significantly under that expected and for which reasons can be sought.

3/45 Not very valuable
6/46 Interesting! Have not really had the time to analyse the data in too much detail
4/47 Not very - fairly irrelevant
7/48 Moderately. The major drawback is that in many groups the samples are so small that it is risky to draw many conclusions from them. Inspectors were impressed by our use of ALIS
3/49 -
4/50 Not much
9/51 Helpful if used wisely and carefully. there are many other factors which need to be taken into consideration: attitude, home background etc.
9/52 Not very - our scores are, on average, what one would expect
7/53 Very particularly re. inspection
1/54 Therefore it has not been valuable
8/55 (Continued) …in general they have confirmed what we knew or suspected anyway
8/56 Not particularly valuable
8/57 Quite useful interesting
8/58 Not very
9/59 Not sure. Many of the comments/attitudes are contradictory
8/60 (Blank)
4/61 Quite valuable - predictable quite often, but better than using raw results. Allows more detailed analysis of performance of students - and staff
8/62 N/A
8/63 Little. it is impossible to deduce what is causing what - which is the cause and which the effect if either.
4/64 Useful - although I was aware of the issues through common sense and because of a number of years teaching experience (More than 5½!)
3/65 Not at all
1/66 -
9/67 It has not really produced any surprises
7/68 (See other comments)
7/69 moderate value
7/70 not especially valuable
9/71 Focus for 2 above - very good. Inaccuracies have been significant, however
7/72 The value added results and the attitude survey results were encouraging

C2-4:(Other comments)

7/68 The ALIS feedback has been very supportive in the sense that students' comments have been positive. Therefore the message seems to be keep going as you are. I found this form very annoying!

C5: (Comments added)

3/4 Most control is with HOD (coded 2)
7/68 ('it' circled)ALIS? not clear

D: Responsibility for students' exam performance: other factors (if any particular ones, please list them)

2/1 F = esp commitment to paid work outside college (no of hours worked)
3/5 F = negative factors - family/personal problems, loss of motivation, etc.
3/7 F = interest
3/9 I have assumed that exam performance means absolute performance - ie a B is better than a C regardless of who scored that particular grade
1/11 For a subject of the nature of physics, I feel that natural ability is by far the most important aspect. This would not be the same for all subjects. Students' home background can have a great effect in individual cases, but in general is less significant.
3/12 Interesting question. Does ALIS have any sane(?) way of 'measuring' home background? It's probably the major factor.
3/15 difficult!

4/17 Possiblyish
4/18 F = if the subject at A or AS is as they expected it to be
3/20 Not easy to do. F could include Dyslexia, Hay fever, other physical problems, special needs etc. Students' 'ability' is not easy to define - not really separable form character attributes. What is a typical student? My answer to D is really saying the students themselves are the most important influence. However, there may be some teachers who are so dreadful their influence is decisive (or even so fantastic).
3/25 F = personal/medical difficulties
6/29 F = in my case, students' mathematical ability (for exam perf in physics)
1/31 F = support
5/33 F = extent to which previous education has prepared them. F' = attendance. I was tempted to add a segment for 'examiners' whims'! There do seem to be discrepancies between one year's results and the next which can be explained no other way
4/34 what is a typical student? It's a different pie graph for each of them and there is more than one category of typical!
4/37 F = Peer group pressure, luck - good or bad. (Circle not divided up) Almost equal given adequate ability
1/39 F = Especially in English - something indefinable - could be as simple as how they're feeling on the day of the exam - ?
1/44 F = Psychological state on day of exam
3/45 I don't believe this is a possible task: obviously the above factors are all important, but the relative importance will depend on individual circumstances. I think home background / class is most important though.
6/46 F = inc. - time doing part time paid work.
7/48 F = previous education (we get them at 16), unforeseen factors (health, financial problems, etc.)
9/52 F1 = resources; F2 = visits abroad (subject is <u>German</u>)
1/54 F = stability / morale in (a) the institution, (b) the education system as a whole
9/59 F = how important a grade in the subject is, eg for university entrance
4/61 F = ?
8/63 D and E: but these may amount to the same factor. You are putting as if they are independent factors. F = class 'buzz' and momentum. (Comments added to pie chart:) C could possibly be bigger and F smaller.
4/64 F = peer and group ethos and pressure
3/65 On average probably about equal, but with any individual student obviously F could make all the difference (or none)
7/69 Assume for A level, F = resources for students and teachers + staff morale + learning environment. NB D: think 'teaching skills' better than 'which teacher'.
7/70 F = detrimental effects: reduced time in class, increased class sizes
9/71 F = 'quality of department'. Difficult qu. (I should have spent more than 30 seconds on it)

E: Please make here any additional comments, including comments on any of the above questions that you found to be unclear, meaningless or otherwise hard to answer:

2/2 It may have been helpful to have been reminded of the variety of analyses which ALIS provides.
3/5 Seemed like a well structured questionnaire to me.
3/7 I personally find ALIS a waste of energy, paper and time. I am in favour of appraisal within the school and an occasional constructive inspection by [subject] teachers.
3/8 I am now dealing with ALIS for this college.
1/10 For physics there appears to be more correlation between GCSE maths and the A level result than between the overall ALIS score.
1/11 I have based my answers in section B on my [main subject] teaching. I feel a lot less confident and less secure about [other subject], which is new to me.
4/18 Comments on questions - please see above. As I have indicated, I do not think ALIS is a sufficiently subtle method of measuring 'value added'. I don't think one exists, or could exist - too many variables. However, I am worried that it may be used as a much more reliable indicator than it really is and that this could be used in eg redundancy choices.

4/19 It would be useful to have the ALIS data on disk instead of printout. The maths data are calculated on a common regression line rather than separate lines for different types of maths. I'd like the analysis to incorporate ethnic identity.

*5/21 Please preserve confidentiality in all the above comments.

4/22 I have a curious interest in the ALIS data, largely owing to the sheer volume of data within the system. However, I feel that there are too many other variables that affect exam performance.

6/27 The questions seem to be loaded against staff. I detect a move to use ALIS in staff appraisal, which will be counter-productive in the end.

6/28 I found this very difficult to fill in as there are so many variables eg q1 (feedback): it depends on the nature of the feedback: objective is hard to quantify; responsibility for students grades - yes of course, but no, at this level too. Hence many answers are down the middle.

4/35 The perceived learning activities section is of no use. The rest of the data is helpful if used judiciously.

4/37 The whole learning / teaching process is very complex and can be affected by all the factors listed above in D. Therefore to reduce it to the questions asked is too simplistic. I am very critical of myself as a teacher and constantly evaluate my work informally. I find this of more use than formal evaluations.

5/38 Could some analysis of how good the regression line predictors are be made? Eg, could it not be that predicted 'A' level maths grade is based just on GCSE maths mark rather than ALIS score? There seem to be large residuals either way in maths and none that near 0 …

1/39 As I have very little knowledge of ALIS so far, I found some of the questions hard to answer. Several of the questions were not wholly relevant for teachers of English, which is a very personal subject, without an exact body of information to 'teach' - results therefore not so obviously obtained during course. (Added at beginning) All my comments / answers relate to my experience teaching English A level - not A level exams in general.

1/40 Have found various questions repetitive. Found C5 doesn't cover my stage - not doing it.

1/41 I would like more information about ALIS. I don't feel I have access to enough information, nor do I know how to use what I have.

1/54 I was worried recently by the article I read (from the Health Service Review?) suggesting that there are statistical flaws in 'league tables' - I did not understand it, but others say it shows that ALIS could be misleading.

8/56 I don't realy know what ALIS is!

8/60 Many of my opinions are based on a current year 13 group who, in spite of my best efforts to entertain them, seem single minded in their determination to be disinterested and to learn as little as possible. I have a sense of impending doom with regard to this year's results and feel very depressed that their failure may be attributed to my personal incompetence.

4/61 Q18 is not clear. Q4 can't be answered by 'disagree strongly' 'agree strongly' etc. Is one agreeing etc. with the belief or with the statement that one is a good teacher.

8/62 Q23: I know nothing about the errors. Q29: I don't know enough about the results to judge.

8/63 Too many questions - you switch off.

4/64 I found some of the questions on ALIS specifically a bit repetitive; hence my answers are rather neutral. I am not fully confident about the extent of its application currently in [institution]

3/65 I am sorry this is so late, Rob. You must be aware that we are all so overburdened now that we are constantly having to prioritise and this got left to the bottom of the pile. I don't really want any of this passed on to anyone at [institution 3]. The whole situation there is too volatile. The pressure is overwhelming and I believe I speak for many others when I say that when I don't teach as well as I'd like to the reason is not that I don't know how to do it better - it's because I'm bloody exhausted! Work overload is the problem, not teaching methods.

7/69 Concerned about lack of mention of resources. Think some questions loaded. Think ALIS could be useful but not if it is going to be used as another weapon to attack teachers.

9/71 Students in maths have been missed out. Classifications (Pure, Further, etc.) have been unworkable. Errors have been regular. Consequently, we have lost considerable faith in the scheme.

	SEX	YRS	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
SEX	1.00	-0.12	0.01	-0.09	0.20	0.12	0.05	0.12	-0.08	0.04	0.05	0.08	0.05	0.30 *	0.07	-0.16	0.11	0.01	0.27 *	-0.06	-0.02	-0.14
YRS	-0.12	1.00	0.07	0.10	-0.14	-0.15	-0.13	-0.06	-0.14	0.01	-0.11	0.04	-0.17	-0.22	-0.23	-0.04	-0.16	0.03	-0.26 *	0.25 *	0.02	-0.01
B01	0.01	0.07	1.00	0.42 **	0.12	0.11	-0.15	0.51 **	0.10	0.35 **	0.05	-0.15	0.27 *	0.21	0.23	-0.10	0.05	-0.13	0.16	0.35 **	-0.18	-0.12
B02	-0.09	0.10	0.42 **	1.00	0.05	-0.01	-0.21	0.27 *	0.23	0.13	0.14	-0.13	0.13	-0.06	0.03	-0.15	-0.20	-0.26 *	0.00	0.21	-0.17	-0.16
B03	0.20	-0.14	0.12	0.05	1.00	0.12	0.14	0.26 *	0.20	0.20	0.08	-0.06	0.26 *	0.52 **	0.22	-0.10	0.01	-0.21	0.62 **	-0.04	-0.07	-0.18
B04	0.12	-0.15	0.11	-0.01	0.12	1.00	-0.26 *	0.08	0.19	-0.07	0.32 **	-0.01	-0.09	0.11	0.05	-0.19	-0.39 **	0.27 *	0.18	0.15	0.00	-0.19
B05	0.05	-0.13	-0.15	-0.21	0.14	-0.26 *	1.00	-0.27 *	0.03	0.07	-0.11	0.13	0.24 *	0.24 *	-0.12	0.25 *	0.26 *	-0.02	0.06	-0.21	0.02	0.16
B06	0.12	-0.06	0.51 **	0.27 *	0.26 *	0.08	-0.27 *	1.00	0.18	0.31 **	0.21	-0.37 **	0.22	0.15	0.47 **	-0.38 **	-0.04	-0.23	0.34 **	0.08	-0.27 *	-0.36 **
B07	-0.08	-0.14	0.10	0.23	0.20	0.19	0.03	0.18	1.00	0.35 **	0.13	-0.02	0.16	0.15	0.04	0.26 *	-0.05	-0.09	0.07	0.03	0.06	0.08
B08	0.04	0.01	0.35 **	0.13	0.20	-0.07	0.07	0.31 **	0.35 **	1.00	-0.02	-0.01	0.19	0.31 **	0.16	0.04	0.02	-0.12	0.26 *	0.14	-0.24 *	-0.07
B09	0.05	-0.11	0.05	0.14	0.08	0.32 **	-0.11	0.21	0.13	-0.02	1.00	-0.47 **	-0.07	0.19	0.27 *	-0.26 *	-0.27 *	0.00	0.19	0.09	-0.39 **	-0.16
B10	0.08	0.04	-0.15	-0.13	-0.06	-0.01	0.13	-0.37 **	-0.02	-0.01	-0.47 **	1.00	-0.08	-0.05	-0.30 *	0.25 *	0.15	-0.09	-0.06	-0.01	0.66 **	0.35 **
B11	0.05	-0.17	0.27 *	0.13	0.26 *	-0.09	0.24 *	0.22	0.16	0.19	-0.07	-0.08	1.00	0.31 **	0.17	0.35 **	0.26 *	-0.08	0.07	0.12	0.01	0.16
B12	0.30 *	-0.22	0.21	-0.06	0.52 **	0.11	0.24 *	0.15	0.15	0.31 **	0.19	-0.05	0.31 **	1.00	0.35 **	0.00	0.13	-0.20	0.60 **	0.11	-0.18	0.05
B13	0.07	-0.23	0.23	0.03	0.22	0.05	-0.12	0.47 **	0.04	0.16	0.27 *	-0.30 *	0.17	0.35 **	1.00	-0.16	0.15	-0.26 *	0.31 **	-0.01	-0.36 **	-0.26 *
B14	-0.16	-0.04	-0.10	-0.15	-0.10	-0.19	0.25 *	-0.38 **	0.26 *	0.04	-0.26 *	0.25 *	0.35 **	0.00	-0.16	1.00	0.33 **	0.24 *	-0.32 **	0.16	0.35 **	0.81 **
B15	0.11	-0.16	0.05	-0.20	0.01	-0.39 **	0.26 *	-0.04	-0.05	0.02	-0.27 *	0.15	0.26 *	0.13	0.15	0.33 **	1.00	-0.15	-0.14	-0.18	0.15	0.30 *
B16	0.01	0.03	-0.13	-0.26 *	-0.21	0.27 *	-0.02	-0.23	-0.09	-0.12	0.00	-0.09	-0.08	-0.20	-0.26 *	0.24 *	-0.15	1.00	-0.18	0.20	0.06	0.24 *
B17	0.27 *	-0.26 *	0.16	0.00	0.62 **	0.18	0.06	0.34 **	0.07	0.26 *	0.19	-0.06	0.07	0.60 **	0.31 **	-0.32 **	-0.14	-0.18	1.00	0.02	-0.27 *	-0.28 *
B18	-0.06	0.25 *	0.35 **	0.21	-0.04	0.15	-0.21	0.08	0.03	0.14	0.09	-0.01	0.12	0.11	-0.01	0.16	-0.18	0.20	0.02	1.00	-0.17	0.14
B19	-0.02	0.02	-0.18	-0.17	-0.07	0.00	0.02	-0.27 *	0.06	-0.24 *	-0.39 **	0.66 **	0.01	-0.18	-0.36 **	0.35 **	0.15	0.06	-0.27 *	-0.17	1.00	0.43 **
B20	-0.14	-0.01	-0.12	-0.16	-0.18	-0.19	0.16	-0.36 **	0.08	-0.07	-0.16	0.35 **	0.16	0.05	-0.26 *	0.81 **	0.30 *	0.24 *	-0.28 *	0.14	0.43 **	1.00
B21	-0.08	-0.22	-0.07	-0.30 *	0.12	-0.15	0.44 **	-0.25 *	0.13	0.02	-0.26 *	0.16	0.43 **	0.24 *	-0.04	0.56 **	0.51 **	0.21	-0.07	-0.11	0.14	0.52 **
B22	0.33 **	-0.41 **	0.05	0.00	0.53 **	0.27 *	0.09	0.20	0.22	0.11	0.14	-0.01	0.16	0.46 **	0.10	-0.13	-0.10	-0.05	0.61 **	-0.03	0.00	-0.10
B23	0.06	0.07	-0.09	0.09	-0.09	0.02	0.08	-0.04	0.01	0.02	-0.14	0.13	0.17	-0.13	-0.39 **	0.13	-0.09	0.21	-0.06	-0.04	0.22	0.12
B24	0.28 *	-0.21	0.24 *	-0.17	0.17	0.00	0.03	0.25 *	-0.20	0.09	0.06	-0.13	0.35 **	0.45 **	0.49 **	0.01	0.13	0.02	0.26 *	-0.04	-0.12	0.09
B25	0.17	-0.18	-0.01	-0.10	0.06	0.73 **	-0.33 **	-0.08	0.00	-0.20	0.20	0.02	-0.15	0.16	0.06	-0.20	-0.43 **	0.32 **	0.28 *	0.20	0.03	-0.22
B26	-0.11	-0.18	-0.38 **	-0.44 **	-0.10	-0.17	0.41 **	-0.49 **	0.03	-0.22	-0.17	0.22	0.09	0.09	-0.17	0.53 **	0.32 **	0.28 *	-0.11	-0.17	0.27 *	0.48 **
B27	0.26 *	-0.06	0.08	0.09	-0.10	-0.01	0.07	0.13	-0.07	0.15	0.06	-0.29 *	0.00	0.04	0.10	-0.12	-0.01	0.10	0.06	0.19	-0.38 **	-0.12
B28	0.19	-0.14	0.16	-0.01	0.17	-0.11	-0.03	0.30 **	0.04	0.15	-0.08	0.03	0.17	0.20	0.20	-0.18	0.31 **	-0.40 **	0.19	-0.07	-0.04	-0.17
B29	-0.13	0.06	-0.22	-0.18	-0.16	-0.19	0.14	-0.37 **	-0.10	-0.06	-0.12	0.28 *	0.08	-0.02	-0.46 **	0.32 **	0.00	0.05	-0.22	0.04	0.28 *	0.35 **
AWARE	0.01	0.28 *	0.14	0.01	-0.16	0.26 *	-0.04	-0.19	-0.27 *	-0.19	0.15	0.11	-0.22	0.00	-0.03	-0.15	-0.07	-0.10	-0.04	-0.03	0.02	-0.04
STAGE	-0.06	0.24	-0.05	0.03	-0.18	-0.25 *	0.11	-0.18	-0.11	-0.15	-0.06	0.20	-0.19	-0.17	-0.16	0.03	0.11	-0.27 *	-0.28 *	-0.32 *	0.32 **	0.16
RESP_ABL	-0.19	0.01	0.06	0.10	0.06	-0.23	-0.04	-0.08	0.01	0.10	-0.06	0.12	0.06	-0.13	0.08	0.10	-0.01	-0.15	-0.01	-0.02	0.05	0.07
RESP_BGD	-0.06	-0.18	-0.06	-0.12	0.00	0.26 *	0.03	-0.01	-0.16	-0.07	-0.01	-0.08	0.07	0.02	-0.01	-0.07	-0.02	0.14	0.04	0.05	-0.22	-0.12
RESP_CHR	0.12	0.00	-0.10	-0.17	0.00	0.22	-0.09	0.02	-0.03	-0.25 *	0.35 **	-0.07	-0.21	0.09	-0.10	-0.08	-0.09	0.09	0.17	0.09	0.14	0.08
RESP_TCH	0.16	0.12	0.08	0.14	-0.13	-0.10	0.13	0.08	0.09	0.07	-0.01	0.01	0.13	0.02	0.02	0.10	0.16	-0.11	-0.17	-0.01	-0.01	0.11
RESP_SCH	0.13	-0.06	-0.02	0.06	-0.02	0.03	-0.03	0.09	-0.05	-0.05	-0.26 *	0.12	-0.10	-0.08	-0.08	-0.08	-0.04	0.08	-0.11	-0.03	0.11	-0.11
RESP_OTH	0.07	0.14	0.02	-0.01	0.03	-0.05	0.10	0.00	0.22	0.20	-0.12	-0.17	0.06	0.20	0.02	-0.03	0.07	0.03	-0.03	-0.09	-0.11	-0.13
MEAN ABS R	0.12	0.13	0.15	0.14	0.15	0.16	0.14	0.21	0.11	0.14	0.15	0.15	0.16	0.18	0.18	0.20	0.16	0.15	0.19	0.11	0.17	0.20
MEAN R SQD	0.021	0.026	0.036	0.030	0.043	0.044	0.031	0.063	0.020	0.028	0.036	0.041	0.036	0.054	0.050	0.069	0.040	0.032	0.062	0.020	0.052	0.065
NUM (R >.3)	1	1	5	3	3	3	3	10	1	4	4	4	4	7	7	9	6	2	6	2	7	7
NUM (R >.4)	0	1	2	2	3	1	2	3	0	0	1	2	1	4	3	3	2	0	3	0	2	4
NUM (R >.5)	0	0	1	0	3	1	0	1	0	0	0	1	0	2	0	3	1	0	3	0	1	2
NUM (R >.6)	0	0	0	0	1	1	0	0	0	0	0	1	0	1	0	1	0	0	3	0	1	1
NUM (R >.7)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

B21	B22	B23	B24	B25	B26	B27	B28	B29	AWARE	STAGE	ABL	BGD	CHR	TCH	SCH	OTH
-0.08	0.33 **	0.06	0.28 *	0.17	-0.11	0.26 *	0.19	-0.13	0.01	-0.06	-0.19	-0.06	0.12	0.16	0.13	0.07
-0.22	-0.41 **	0.07	-0.21	-0.18	-0.18	-0.06	-0.14	0.06	0.28 *	0.24	0.01	-0.18	0.00	0.12	-0.06	0.14
-0.07	0.05	-0.09	0.24 *	-0.01	-0.38 **	0.08	0.16	-0.22	0.14	-0.05	0.06	-0.06	-0.10	0.08	-0.02	0.02
-0.30 *	0.00	0.09	-0.17	-0.10	-0.44 **	0.09	-0.01	-0.18	0.01	0.03	0.10	-0.12	-0.17	0.14	0.06	-0.01
0.12	0.53 **	-0.09	0.17	0.06	-0.10	-0.10	0.17	-0.16	-0.16	-0.18	0.06	0.00	0.00	-0.13	-0.02	0.03
-0.15	0.27 *	0.02	0.00	0.73 **	-0.17	-0.01	-0.11	-0.19	0.26 *	-0.25 *	-0.23	0.26 *	0.22	-0.10	0.03	-0.05
0.44 **	0.09	0.08	0.03	-0.33 **	0.41 **	0.07	-0.03	0.14	-0.04	0.11	-0.04	0.03	-0.09	0.13	-0.03	0.10
-0.25 *	0.20	-0.04	0.25 *	-0.08	-0.49 **	0.13	0.30 **	-0.37 **	-0.19	-0.18	-0.08	-0.01	0.02	0.08	0.09	0.00
0.13	0.22	0.01	-0.20	0.00	0.03	-0.07	0.04	-0.10	-0.27 *	-0.11	0.01	-0.16	-0.03	0.09	-0.05	0.22
0.02	0.11	0.02	0.09	-0.20	-0.22	0.15	0.15	-0.06	-0.19	-0.15	0.10	-0.07	-0.25 *	0.07	-0.05	0.20
-0.26 *	0.14	-0.14	0.06	0.20	-0.17	0.06	-0.08	-0.12	0.15	-0.06	-0.06	-0.01	0.35 **	-0.01	-0.26 *	-0.12
0.16	-0.01	0.13	-0.13	0.02	0.22	-0.29 *	0.03	0.28 *	0.11	0.20	0.12	-0.08	-0.07	0.01	0.12	-0.17
0.43 **	0.16	0.17	0.35 **	-0.15	0.09	0.00	0.17	0.08	-0.22	-0.19	0.06	0.07	-0.21	0.13	-0.10	0.06
0.24 *	0.46 **	-0.13	0.45 **	0.16	0.09	0.04	0.20	-0.02	0.00	-0.17	-0.13	0.02	0.09	0.02	-0.08	0.20
-0.04	0.10	-0.39 **	0.49 **	0.06	-0.17	0.10	0.20	-0.46 **	-0.03	-0.16	0.08	-0.01	-0.10	0.02	-0.08	0.02
0.56 **	-0.13	0.13	0.01	-0.20	0.53 **	-0.12	-0.18	0.32 **	-0.15	0.03	0.10	-0.07	-0.08	0.10	-0.08	-0.03
0.51 **	-0.10	-0.09	0.13	-0.43 **	0.32 **	-0.01	0.31 **	0.00	-0.07	0.11	-0.01	-0.02	-0.09	0.16	-0.04	0.07
0.21	-0.05	0.21	0.02	0.32 **	0.28 *	0.10	-0.40 **	0.05	-0.10	-0.27 *	-0.15	0.14	0.09	-0.11	0.08	0.03
-0.07	0.61 **	-0.06	0.26 *	0.28 *	-0.11	0.06	0.19	-0.22	-0.04	-0.28 *	-0.01	0.04	0.17	-0.17	-0.11	-0.03
-0.11	-0.03	-0.04	-0.04	0.20	-0.17	0.19	-0.07	0.04	-0.03	-0.32 *	-0.02	0.05	0.09	-0.01	-0.03	-0.09
0.14	0.00	0.22	-0.12	0.03	0.27 *	-0.38 **	-0.04	0.28 *	0.02	0.32 **	0.05	-0.22	0.14	-0.01	0.11	-0.11
0.52 **	-0.10	0.12	0.09	-0.22	0.48 **	-0.12	-0.17	0.35 **	-0.04	0.16	0.07	-0.12	0.08	0.11	-0.11	-0.13
1.00	0.07	0.02	0.19	-0.27 *	0.59 **	-0.05	0.18	0.19	-0.14	0.01	0.10	0.00	-0.16	0.07	-0.11	0.02
0.07	1.00	-0.09	0.29 *	0.31 **	-0.03	0.12	0.17	-0.15	-0.17	-0.19	-0.08	0.02	0.32 **	-0.24	-0.15	0.07
0.02	-0.09	1.00	-0.19	0.00	-0.05	-0.02	-0.20	0.33 **	-0.12	0.07	-0.01	-0.02	-0.06	0.19	-0.06	-0.04
0.19	0.29 *	-0.19	1.00	0.06	0.00	0.06	0.10	-0.14	-0.07	-0.07	-0.07	0.01	0.16	-0.11	-0.01	0.06
-0.27 *	0.31 **	0.00	0.06	1.00	-0.15	-0.06	-0.14	-0.23	0.19	-0.20	-0.23	0.22	0.18	-0.18	0.04	0.14
0.59 **	-0.03	-0.05	0.00	-0.15	1.00	-0.18	-0.11	0.36 **	-0.06	0.00	0.03	0.02	0.05	-0.04	-0.08	-0.04
-0.05	0.12	-0.02	0.06	-0.06	-0.18	1.00	0.26 *	-0.12	0.05	-0.17	-0.25 *	0.05	0.13	0.18	0.04	0.13
0.18	0.17	-0.20	0.10	-0.14	-0.11	0.26 *	1.00	-0.21	0.06	-0.05	0.00	-0.22	0.00	0.19	0.01	0.09
0.19	-0.15	0.33 **	-0.14	-0.23	0.36 **	-0.12	-0.21	1.00	0.03	0.29 *	-0.03	0.12	-0.04	0.05	0.02	-0.10
-0.14	-0.17	-0.12	-0.07	0.19	-0.06	0.05	0.06	0.03	1.00	0.38 **	-0.01	-0.01	0.01	0.16	-0.05	-0.08
0.01	-0.19	0.07	-0.07	-0.20	0.00	-0.17	-0.05	0.29 *	0.38 **	1.00	-0.12	-0.07	-0.01	0.30 *	0.00	0.04
0.10	-0.08	-0.01	-0.07	-0.23	0.03	-0.25 *	0.00	-0.03	-0.01	-0.12	1.00	-0.54 **	-0.34 **	-0.36 **	-0.40 **	-0.43 **
0.00	0.02	-0.02	0.01	0.22	0.02	0.05	-0.22	0.12	-0.01	-0.07	-0.54 **	1.00	-0.10	0.10	0.05	-0.05
-0.16	0.32 **	-0.06	0.16	0.18	0.05	0.13	0.00	-0.04	0.01	-0.01	-0.34 **	-0.10	1.00	-0.19	-0.17	-0.16
0.07	-0.24	0.19	-0.11	-0.18	-0.04	0.18	0.19	0.05	0.16	0.30 *	-0.36 **	0.10	-0.19	1.00	-0.03	0.03
-0.11	-0.15	-0.06	-0.01	0.04	-0.08	0.04	0.01	0.02	-0.05	0.00	-0.40 **	0.05	-0.17	-0.03	1.00	0.28 *
0.02	0.07	-0.04	0.06	0.14	-0.04	0.13	0.09	-0.10	-0.08	0.04	-0.43 **	-0.05	-0.16	0.03	0.28 *	1.00
0.19	0.17	0.10	0.14	0.18	0.19	0.11	0.14	0.16	0.11	0.15	0.12	0.09	0.12	0.11	0.09	0.09
0.062	0.051	0.018	0.034	0.050	0.062	0.020	0.028	0.041	0.020	0.032	0.033	0.018	0.023	0.020	0.014	0.016
7	7	2	3	5	9	1	3	6	1	4	5	1	3	2	1	1
6	4	0	2	2	6	0	0	1	0	0	3	1	0	0	1	1
4	2	0	0	1	2	0	0	0	0	0	1	1	0	0	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

	less than 5 mins	5-20 mins	20mins - 1hr	more than 1hr	missing
Time spent on 'TARGETS 97'	5	9	1	0	0
Time spent on 'RESULTS', 'GRAPHS' for 94-6	4	8	3	0	0
Time expected to spend on 'TARGETS 97'	8	5	2	0	0
Time expected to spend on 'RESULTS', 'GRAPHS' for 94-6	7	5	3	0	0

	Yes	No	missing
Discussed with colleagues?	11	3	1

	very easy	easy	mod. hard	hard	imposs.	missing
How easy to understand?	3	7	1	1	2	1

	extremely useful	useful	of some use	no use at all	missing
Usefulness of 'Targets 97'	1	1	10	2	1
Usefulness of 'Student Results'	1	6	7	0	1
Usefulness of 'Graphs'	1	5	8	0	1
Usefulness of 'Class Averages'	0	5	8	1	1
Usefulness of 'Summary by Teacher'	1	6	7	0	1

1 Can't believe some of results for weak students, some predicted C will be lucky to pass. Suspect maths GCSE result much more relevant to A level results. Sorry, I'm not a believer.

2 We had done these analyses ourselves.

5 Note: the time I have spent considering the data must be considered in the context of only having eight students worth of data.

7 I have found it very difficult to understand the charts and analysis – could it be presented in a more user-friendly form?

8 The 'targets 97' material is if little practical use since:
- it came too late really in the academic year to be of much value in student review/action planning sessions;
- the minimum grade suggested is not as good a guide as our knowledge of the students potential and capabilities based on 18 months of working with them;
- many of us (staff) are dubious of working with ALIS in this way to individual students, rather than got identifying patterns and changes in groups of students.
The other second set of information was useful and, indeed, interesting. I have to admit not having had sufficient time yet to digest it properly and fully. As head of department, I would actually have liked to see the information on each of the teachers - although I can ask them for it. We generally analyse our results in Autumn, and this information would be tremendously useful at that time of year, rather than now.

9 Is this done anywhere else in the world? USA? It is useful, but I don't think it is essential.

11 Re Q.5: We don't think many of the statistics are clearly enough presented for most non-statisticians to interpret easily. For example, the ideas of significance and deviation are not simple for non-mathematicians and indeed the significance values quted in the tables seemed of doubtful value in any case. THEREFORE, we suggest you summarise and simplify the stats which you think are most useful and add a health warning if necessary.
7: Is any account taken of students' absence? They may have achieved highly at GCSE, especially from smaller independent schools with pushing, etc. and full attendance, but aren't so able in large 6th form environment. Two students who, according to your statistics, significantly underachieved in fact missed many lessons, despite best effort of senior tutors to pursue.

13 Questions 5, 6 require some study of the information before they can be answered. We have increased teaching hours (and therefore more preparation) together with much more administrative work these days. The opinion of some staff is that, in order to keep workloads to a tolerable level, less homework and shorter assessments will be set in future. Moreover, increased class sizes (due to 'productivity' requirements) can only have a detrimental effect on results. Class contact time, per A level subject, is currently 5½ hours per week; Twenty three years ago this figure was 7 hours. Staff morale is at an all-time low because of the dispute regarding new contracts: staff who have refused to sign have received no salary increase whatsoever over the last four years, and have been informed that they will, in future, remain on their current salary. I feel that, until at least some of the above situations a reversed, there can be no improvements in overall teaching, and that ALIS will not have the application as intended.

Please describe any changes which may have resulted from your involvement in this project, specifically:

('f' denotes those who had received feedback)

1. Changes in your attitude towards ALIS and the feedback it provides:

1f A greater awareness of what ALIS can provide
2 Now take it more seriously
3f A greater awareness of pupil achievement/attainment
5 I don't think my attitude to ALIS has changed because of my involverment in this project. I had quite a positive attitude to it before. Analysing students' performance on the basis of value added is fairer than looking at raw results. The new government seems to be taking this on board too.
6 More positive
7 I find it less threatening than I used to because evaluation has become feature of many spheres of life.
8 'Perceived Learning Activities' not very helpful, as they are so generalised as to be difficult for students to understand and difficult to apply to specific subjects. Better to assess this aspect internally.
9 I haven't really studied ALIS feedback as an individual subject teacher
10f None
11f None
12f No
15f Although info received is of use, in my case where group sizes are very small the generalisations of the data do not always fit specific cases. This is a problem in terms of using actual data.
16f Not much – except to deepen my interest in the use of MPGs if the information is received/used early enough. Also t realise how useful comparative data (between subject performance for a student, say) can be.
17f I am getting used to it!
18 (1) Less wary of using data during self assessment: I can see the value in it increasingly.
(2) Feel more confident about analysing the data and am more aware of its potential in student assessment programmes too
(3) Imp. tool for college tutorial meetings with individual students, esp when your discussing subjects you don't teach them
19 Always interested to read information on the subject. Still not entirely clear about how ALIS is used by institutions
20 The key issue in the residuals:- a consistent pattern emerging over a longitudinal period does seem a fair indication of the effectiveness of reading. Surveys of student attitude should be seen in this context (i.e., if a teacher had consistent value added success and <u>good retention</u> – something that ALIS doesn't pick up – negative attitudes perhaps don't carry a great deal of weight.
21 I am finding out more about it
22 Easier to understanding after a few years
24 Your project happens to coincide with an inset day about student target grades. So I have changed my mind, but no single event has been responsible rather an accumulation of directed thought!
25 Although I am aware of ALIS, I have not been specifically involved in feedback about it. My involvement has been minimal.
26 None. It has its place giving a guide on expected performance from previous results, but other factors such as motivation and work effort cannot be measured by ALIS and can change and distort expectations.
27 None
28 Unaware of any
30f No change – I still feel generally doubtful about the entire process!
32f I haven't yet had time to analyse the feedback sufficiently – I find the presentation of the data difficult to understand and put to effective use.

33 I still feel very ambivalent about ALIS. GCSE grades depend on so many factors, eg home background, peer group, area, social class, quality of teaching. A level requires more natural flair and ability
34 In favour of ALIS principles for evaluating progress, but still unsure of value of additional information provided by using an outside agency.
35f None
36f I have always had a positive attitude towards ALIS and the quantity and quality of the feedback has improved significantly over the years. I am hoping that YELLIS will also be taken up by our contributory schools as we would find this information very useful.
37f This questionnaire was a bit remote for me as the class that took A level French was in 95.
38f I have been more sensitive to the ALIS scores of incoming students. However, I have also been surprised that their performance at A level English often fails to match these scores.
39f Seems too complicated, and I have strong doubts about its ability to distinguish between students who have similar 'scores' for very different reasons. Does not take into account the candidates conscientiousness as students.
40f Find it useful but value added predictions indicate alarmingly a mismatch between GCSE and A level.

2. Any changes in how you will use ALIS feedback in the future:

1f More detailed examination and evaluation of 'trends' as no. of years' data increases
2 Not sure
3f Greater discussion/dissemination amongst colleagues. Staff need to be encouraged to use such data to reflect back on their teaching – so long as it is not used against them!!
5 This year we did not get the data on students' perceptions of teaching and learning styles in our institution. This is one of the most useful aspects of the data, especially regarding q.3, below
7 I will think about using is as a basis for target setting, though this would be best carried out as part of a whole college policy.
8 In predicting expected grades early on to set targets for students.
9 I may pay more attention to it
10f None
11f None
12f No
16f I will think about using MPGs in review and targeting sessions with students. I will analyse comparative data (as HoD) more if it is available easily
17f Useful for my stats students to do mini-projects
18 More for self appraisal in line with quality assurance evidence for our department. To help spot students who may not be achieving their full potential. As above (3) in my role as personal tutor rather than subject tutor.
19 For internal discussion within the section
20 As you can see from my answers to questionnaire O have been convinced by my involvement with ALIS of the validity of the information it presents. As a manager, I would find it very useful in monitoring the performance of my team. As a teacher I can understand why people feel threatened by it
21 We are going to set target minimum grades
22 No
24 I have started, with some students, to talk in terms of the standards they should always be aiming at. Trying to help them be confident and determined enough to reach their target.
25 N/A
26 use only as a guideline to how well a student might do.
27 Aim to give each student a target grade, although these could be fairly depressing!
28f Unlikely
32f Unknown
33 No

34 ALIS prediction useful for individual students at enrolment and when experiencing difficulties. Intend to conduct analysis of students dropping subject based on GCSE point scores
35f None
36f I hope we 'milk it' effectively now. It informs our action planning and departmental plans and is useful when reviewing/interviewing individual students.
38f No
39f No: will still analyse the results for individual sets and hope that the results of this analysis match my/our own knowledge (perception) of the reasons for these results.

3. Any changes in your teaching:

2 Will change and see results if ALIS feedback suggests
3f To try different strategies in particular to accommodate the gender difference
5 Not as a result of this project
7 More aware of meeting students' needs rather than just getting through the syllabuses
9 Don't see how it would affect my teaching method at present
10f None
11f None
12f No
14 Not yet
16f Not really
17f Not really
18 Reflecting on changes in T & L styles currently. New plans for induction programme to begin Sept 97.
19 Have not yet received information for consideration – so not sure what you mean here.
20 Not yet
21 None so far, that I have detected, but there may be some in the future.
22 I'm always changing that – regardless of ALIS.
24 Changes in subject tutoring, not subject teaching.
25 N/A
26 No
27 Aim to chase up those students who fall by the wayside, with bad attendance and low work-rate, etc. as they are the ones with the massive negative residuals.
28f Unlikely
32 None to date
33 Maybe
34 Not really
35f None
36f It does focus the mind on individual progress rather than the performance of the group as a whole and this is a positive outcome, and does affect the way you teach. It has encouraged an ongoing debate about teaching and learning styles which is very healthy and which has also had an impact on classroom techniques.
38f No
39f Yes, but arising from staffing constraints rather than from what ALIS has told me (answer should therefore probably be No!?)
40f Am trying to analyse ways in which I can fill gaps in knowledge and spelling and grammar which I used to take for granted.

Do you have any other ideas about how the feedback from ALIS could be improved?

5 Feedback on AS subjects has not always been forthcoming or as full as for A level. Feedback could arrive more promptly – and in full.
9 No!
10f No time to think about it.
11f No
12f The feedback is OK in parts, although many errors in how the analysis is done, but there are lots of mistakes in the material put out prior to the analysis. We could talk on the phone about this if you want sometime.

14	I still don't believe that I receive information in a form which focuses me on potential changes/improvements I could make.
16f	I would like minimum predicted grade information early in first year; and ALIS data to include ethnic category information for early analysis.
17f	Can't there be some feedback from the individual subject GCSE results – perhaps weighting results.
24	Is ALIS a reliable indicator for all subjects and to the same degree?
25	N/A
27	No
28f	Students' attendance and commitment (perhaps as measured by meeting of deadlines, etc) could also be taken into account.
29	Faster service
30f	See my comments on earlier return – too complex for non-statisticians!!
33	No
36f	If any incoming students have a YELLIS profile this would be useful to have. A single list of residuals in all subjects for an individual student.
38f	I found the presentation of information rather confusing. Too much graphical communication and numerical bias. In the end I just gave up trying to understand it.

A (chosen for rating feedback 'very easy' to understand)

I: There are a few specific questions that I want to ask, just to check whether my interpretations of some of the things on the questionnaire are right. These are asking you to rate on a scale from 0 to 10, where 5 is in the middle. So, first question, then: To what extent do you perceive your students' success or failure as being within your control, where 0 would be 'nothing to do with me' and 10 is 'totally in my control'?

A: I would put that probably about 7 or 8.

I: OK, and do you think that your feeling about that has changed over say the last year?

A: It's changed in as much as the majority of physics students do modular physics and modular physics allows them a lot more control over how well they do than the traditional physics where you work for two years and do an exam and what we find is not necessarily just the better ones but the ones that have become more keen, and they become more ambitious, they can resit modules and of course by resitting modules you can encourage them to perform better. So, by and large people get better and better. There are exceptions to that of course, but under our constant driving but that's what it is all the time. It's constant driving. It's very stressful.

I: That's interesting. So none of that has anything to do with ALIS at all?

A: No. What we find, the most interesting thing with ALIS is our raw results are slightly above the national average – we ended up this year with 100% pass and our percentage pass at every grade was higher than the national, apart from the grade A where nationally it's 21% and we had 19%. But when you consider the nature of the people that we have here, I expect to see that our ALIS results should be very positive.

I: Indeed. Next question then: How confident do you personally feel about your effectiveness as a teacher? So 0 would be not confident at all …

A: At the moment probably about 9 or 10, although I'm becoming incredibly disillusioned and feel very close to packing it all in, but we'll say 9 or 10.

I: OK, and has that changed, say over the last year?

A: Yes, but again that's nothing to do with ALIS, that's because of conditions in this college.

I: OK, well that's useful to know anyway. So to what extent, talking about ALIS then, do you believe the information they provide, in other words the value added, to be a fair measure of performance?

A: I think I would put 10 at that. 9 or 10 anyway.

I: So you have quite a lot of faith in it?

A: Yes, with big numbers. I don't have a lot of faith with small numbers. You could isolate a few students – that's the thing about ALIS – you've got this whole spectrum of attitudes. It's dead easy, if you could just throw out a few students you could have even more spectacular ALIS results.

I: And again has that changed, do you think?

A: No, I think that's just the human race, basically.

I: And generally, your attitude to ALIS, where 0 was very negative to 10 very positive, could you put a number on it?

A: I would put it about 9 on that.

I: OK. I have looked at the pattern of results in your department, and you are one of the best, in terms of residuals, in the country.

A: Really?

I: I wondered whether you knew that? You sound surprised?

A: Well we've noticed when we see the lists that [name of institution] comes out pretty near the top on residuals, but it just reflects the huge effort we put into students here. It's not all due to the course, the linear course still does very well, and there are definitely problems with the modular course – this year three people missed a grade A by the skin of their teeth, so we only got 7 grade 'A's instead of 10, which would have made the ALIS results even more spectacular than they are.

I: OK. Thank you very much.

A: The thing about ALIS, by the way, is that the actual raw residuals and hierarchy of residuals are very interesting. What I find really boring is all the stuff about attitudes. This is another thing we have to do in this college, we have to give them initial course questionnaires, mid-course questionnaires, end-of-course questionnaires, and certainly the

initial and mid-course questionnaires are a waste of time. My analogy was it's like asking a patient what they think of difficult medicine – very poorly tasting medicine – half way through a course. The time to ask the patient is at the end of their convalescence, how did they find their course of treatment, if they survived it, as it were. And physics is not an easy subject to get into their heads, you have to drive them on all the time, at least with the kind of students we get here. You get a hell of a wide spectrum of attitudes – not even abilities, but attitudes – they just don't want to do any work. So we have to instil in them the idea that they can succeed and do well.

B (chosen for big changes on questionnaire constructs)

(First part of interview not recorded audibly)

B: … and now that it's modular – I've often thought you ought to explore this – we get from the board an individual printout of each module, and often of course, one of the modules is much lower than the other three, and that sort of thing. If you had those individual modular marks you could work out the individual teacher's value added much more precisely than just by saying they shared a set.

I: Because each teacher teaches a separate module?

B: Yes, everybody teaches a separate thing. I don't want to encourage this because obviously it's a bit …

I: A bit sensitive?

B: I don't want everybody thinking if I do a module badly then I can't be any good, or whatever. But it would actually be possible in the long run, wouldn't it, to do that?

I: Absolutely, yes.

B: And it would be more accurate and it might even be easier for you, mightn't it, than messing around with who shared sets and so on?

I: Yes, it might well be. Certainly, there are a number of modular subjects now, it's quite a growth area, and where it is the case that different teachers teach different bits of it, it is quite possible … Actually, your comment about it being sensitive, one of the questions I asked was whether people thought that information should go to the individual teacher or the head of department or the ALIS coordinator. Most people seemed to be quite happy about that …

B: I think we are because we are more or less confident. I don't know what my colleagues in the department said, but I'm sure I put that, didn't I?

I: Yes, you did. When you say 'confident', do you mean confident with ALIS?

B: I'm not sure about that! I think we're confident that we're not expecting to be caught out as not very good teachers, perhaps hubristically, but I don't think anybody is really terrified that they're going to be exposed as the weak link or anything like that. As far as general confidence about ALIS is concerned, it just seems … in ways I can't quite put my finger on, it seems rather irrelevant to the results the students actually get, taking into account all sorts of factors that ALIS can't take into account, such as whether they worked terribly hard or whatever. In this years' results, which on percentage terms were pretty strong, I had lots of people who'd skived and didn't work and ended up with Bs and Cs, as well as the ones you'd expect them to get. Suppose they had good GCSE scores, and they had worked for a bit and then gone off the boil and skived and so on, and had ended up with a B, it would be quite difficult, wouldn't it, to tell quite where the responsibility lay for them getting a good result, let alone a bad one?

I: Yes, indeed. In fact that's one of the things that I'm hoping to try an find out: 'to what extent is the teacher responsible for the performance of the students?', because nobody knows. There is all this talk about 'teacher effectiveness' and schools having to set targets and so on, and it may be that the teacher can't really do very much about it. We don't know that.

B: Yes, I think so …

I: You think they can't?

B: I had three sets last year … it's incredibly complex isn't it as soon as you start thinking about it? One group I took over from two members of staff who left, so both of their teachers left. I and a probationary teacher who had just joined the college staff took them over. They were the usual mixture of hard working and not very hard working – within

the same person sometimes! And in the end their results were really pretty good. I had another set who were totally boring and their results were boringly predictably pretty good, and then I had another set who had always been not very good on attendance and several people left it and illness and all sorts of psychological traumas – real awful things – and their results on the whole I felt were a bit disappointing, even taking that into account. So that's me, the common link between those three sets, with a great variety of people I was sharing with too – experienced and probationary. When you throw in all those other things that ALIS looks at – social background and all that sort of stuff – you do end up wondering what it's telling you that's of any use.

I: That's an interesting question. How much of a loss would it be if you never had anything from ALIS again?

B: It is interesting, but in another way it would be quite a relief really. It is quite complex, especially for people like me who are not particularly statistics-minded, actually sitting down and doing it when you wish you could be preparing Jane Austen or something, but I have always had the feeling that one ought to be using it in some very sophisticated and proactive way … If I was really on the ball with this stuff and I looked at all those graphs, I really could go and see where all the problems were – where my own individual problems were – and do things differently. Without it I just wouldn't feel that sense of guilt. I suppose I would console myself by saying something fairly bland like 'it didn't really tell me much that I didn't feel I already knew.' Another thing I've always thought about it was that in the first year we had a positive residual of 0.5 or something like that, which is high, isn't it?

I: Yes.

B: And since then – this year it's 0.1, and it's been -0.1 – it's always been around 'results as expected', and I've always thought that I'm sure we didn't do any better in that first year. I'm sure we haven't slipped or made a greater effort then. From that point of view, I think is there some respect in which it's not reliable, is there some way it could be made more reliable?

I: That's another question I'm hoping to get some evidence about. One of the things you said there about being on the ball, I was wondering to what extent you find the whole thing accessible? You said you were not a statistician: is that an obstacle to you? Do you find the numbers …

B: I feel very confident with the very elementary bits of it. I look at the overall residual and breathe a sigh of relief probably, though I do sometimes think, 'are we really not better than expected?' A thing that's often frustrated me is this thing about – of course I can't explain it, but you'll know what I mean – if somebody gets a grade B and their GCSE score recommended that they get a high B, then they're always going to appear with a negative residual, because you can't get a 'high B' – which, by the way, would be one of the advantages of looking at the modular marks, because then you do get the high and the low and the middle. So I'm quite interested in and happy with that. I work out the individual – I haven't yet done it this year – but I will work out the individual averages for teaching groups. We do talk about the people who have achieved significantly, or even wildly, below, and indeed above. Of course the only difficulty with that is that you can always find explanations, can't you? But there have seemed to be genuine explanations as to why they did underachieve. And beyond that, to be honest, I don't go. I look at the distribution graph of people above and below the line and see that they more or less balance out – that sort of thing. So that's really all I do. And this year – you sending us our individual scores as teachers – we looked at them and thought 'that's OK then' or if it wasn't quite – well to be honest I don't know whether it wasn't quite in anybody's cases – I'm sure what people would say is 'Oh well, that's ALIS for you' – there is a certain element of that, obviously.

I: Some of the comments were that people found it a bit incomprehensible. I don't know whether you felt that?

B: As a Curriculum Manager, I've from the beginning been to meetings, I've read the booklet, and every year I refresh my memory about it and will in a while go back and look at it, but yes, I think for your normal busy classroom teacher … It's funny, it's as if the degree of simplicity which would actually be useful would probably be so bland – and it wouldn't be of any interest to you who are dedicating your whole lives to it up there. That's a real problem isn't it?

I: It is a problem, but then it is the users really who should dictate, I think.

B: Yes. What is the position with the government publication of league tables and so on. Are they really going to use this?

I: Value added?

B: Yes.

I: They say they are, yes. I haven't seen the time scale for it yet, but as I understand it, value added is going to become publicly available knowledge, and published in the form of league tables – which of course ALIS has always been against. They've never wanted league tables of any kind.

B: No, because you see your role as supportive, don't you?

I: That's right.

B: That's another thing. Every summer there's this big hoo-ha about results. Of course, now that we're in F.E. we don't tend to appear, they seem to ignore us, I mean it's all schools … but there's no doubt at the beginning of term you know when [the Principal] says, 'Well thanks everybody, great results…' It's all in terms of that and it doesn't make much sense. No, I'll re-phrase that. It doesn't strike home for us, I think, that we've got a good, or whatever, value-added score because in society at large, everybody is saying 'Winchester is wonderful because a hundred percent get grade A' and so on. So, I think if it had a bigger degree of public prominence from that point of view it would be rather good, although obviously, I can quite see why you don't want it, and as soon as it gets into the hands of newspapers and politicians, what chance has it got?

I: Well, yes it's a bit of a difficult one that

B: So, what is the answer to your question? As it is at the moment, it seems, my gut feeling is that it seems a bit more complex than I want it to be, whether it needs to be, and it seems slightly tangential to the main source of pleasure, which is, 'Oh great, 35 people got A's and that was 20 per cent of the year and that's better than last year' – that sort of thing. We still seem to think of it in those terms

I: I've just got four very quick questions…

B: O.K., well fire away

I: These are all rating on a scale of 0 to 10, so 5 is the middle value. To what extent do you perceive your students success or failure as being within your control, so 0 would be nothing to do with me and 10 is totally in my control

B: This was one of the questions on one of your most recent questionnaires wasn't it?

I: Similar, yes.

B: I think I must have said something in the middle, I think I must have said 5.

I: OK, can you say any more about that? Does it mean anything?

B: Well I think it's…. The problem is your students as a sort of group because I've sometimes felt, well, I think every teacher has, 'Yes I've really made a difference there,' whether vanity comes into it or not. Obviously, I'm talking about to the good. I suppose one is less likely, I don't think I've ever felt, not in recent years anyway, 'Oh God I was terrible that ruined their life chances,' but it's so different in the case… For example, yesterday, a boy came from this group I was telling you about, you know the one I took over from [name of teacher] just for the second year of the course and this kid, he'd hardly done an essay. It was terrible trying to wring essays and literary appreciations out of this bloke, he was often absent, I mean I should think well, it always seems worse in memory, but I should think that he'd missed a good quarter of the lessons… And he was obviously naturally quite bright, he had a lot of problems at home, stress and strain and blah, blah, blah and I was thinking, 'Oh well, he'll probably get a D if he's lucky,' and lo and behold, he gets a B. He came in yesterday to say 'I came to say thank you' and I really felt 'What are you thanking me for?' And apart from the fact that every now and again I chivvied you and when you were there I did my best and so on but simply in terms of hours of contact it couldn't have made all that much difference.' He must have done a hell of a lot at the last minute or been incredibly lucky or whatever. I mean, for example, that was the set that was inspected and I did an inspection lesson, which was therefore a bit better than usual, and what we did was the passage that came up in the exam. For once, the whole class was there so everything sort of went to blow him on a gale of success. So, the answer is it's so complex that I don't think one could say one's pupils as a whole……

I: How confident do you feel about your personal effectiveness as a teacher, so 0 would be not confident, 10 extremely confident.

B: Well, again it's in different ways isn't it? I mean, everybody feels less confident than they probably really are deep down I think. But, I think it's true isn't it that the bashing we get in the public eye does actually rob people of confidence. Of course, it's actually also a widespread feeling in education now that because you're being asked to do many things which seem unimportant and merely administrative and nonsensical, like having inspections for example, which personally I feel didn't do anything for us….. You don't feel confident that the job you're doing is being done as well as it could be even though your confident you're doing it as well as you could. I think there's quite a split nowadays in teachers' view of themselves as a person and a classroom teacher and themselves as a cog in this rumbling, complex machine that we've all become. Sorry these aren't clear answers.

I: No, no, it's useful

B: So again it's sort of five isn't it really?!! It depends what you mean by mean, as they used to say on the Brains Trust!!

I: I think what you're saying really is that you can't translate it into a number isn't it?

B: Yes, exactly.

I: Anyway, a couple of questions about ALIS. To what extent would you say that the information provided is a fair measure of performance? Again from 0 to 10, 0 is totally unfair, 10 is entirely fair.

B: I think it's objective isn't it? That's the thing I've always thought about it. I mean, when I've seen somebody who I've taught who got a residual, somebody who for example got a O level – or GCSE rather – score which was 6.65 or something, really quite good, a couple of years ago. Yet when he came here he never really worked very effectively and we didn't, despite our best efforts, manage to bring him up, and he ended up with a negative residual. I thought, not 'Oh that's told me where I've been going wrong,' necessarily, but I think I thought 'Yes, that is a good objective view of it,' and I think it's stopped me saying, 'Oh well [name of student] just never worked it was all his fault.' I think from that point of view, it is valuable. One does feel that it is objective and it is valuable for its objectivity. Though, of course, on the other side there's always this idea that the information you actually get as a whole has got so many variables lying behind it that it's difficult to be sure that the overall picture is telling you much. But as a sort of, not warning light, but making you think about it. That's it isn't it, the big advantage of ALIS is that it's made everybody think about it, whether or not they've been satisfied.

I: Well, yes, obviously there are dangers with the interpretations. It can certainly be over interpreted. And finally then, just a general attitude towards ALIS, would you say that was positive or negative? Can you say that or not?

B: Well, it's certainly not negative. I mean I don't hate the sight of it and wish it would collapse….. Because I feel there is something there that is a good idea and I'm very impressed by the way it's adapted and changed, refined its technique and so on. But, I don't feel strongly positive to be honest. I think it's a sort of worthy enterprise, which is probably…. I think up there you're probably working towards things which are really good and you're obviously there, or a lot are there. If that's of some help – sounds a bit woolly, doesn't it?

I: No, that's extremely helpful, very much so. And I appreciate having had your time.

B: Well I'm sorry I didn't give you the few minutes earlier in the last year, that it would have taken me to get things back to you on time

I: That's OK. Don't worry

B: That was bad I'm afraid…….

C (chosen for large increase in 'ALIS fairness' – control group)

I: You notice that some of the questions were the same questions on the two questionnaires that you had?

C: Yes.

I: And you also know that some of the people have had the feedback that I sent, or perhaps you didn't know that?

C: No.
I: That you hadn't had it? I've sent out class by class analysis of all the groups that you've taught and I'll shortly be sending that to you and all the other people. Some people had that before, in between the two questionnaires you see. The idea was to see whether they'd changed and one of the reasons that I wanted to talk to you was that you didn't have the feedback, but you do seem to have changed! [C laughs] So I wondered whether that was a genuine change or whether it was just a kind of … whether it didn't mean anything.
C: I suspect that it might not have meant anything
I: Right
C: It depends how you were feeling at the time
I: Right, well, if that's the case it's important to know that, obviously
C: Is this taking everybody's views into account or mine personally?
I: Sorry, is what?
C: All of the department?
I: No, just you personally.
C: Right, because the others don't always see so much of the ALIS stuff as I do, so I'm possibly more aware of what it involves.
I: No that's right. I mean one of the things for instance – now what was it I thought about that? Oh yes, general attitudes towards ALIS, particularly in terms of how valid or fair it is, was one of things that I was trying to get at, and on my measure of that you seem to have increased. You seem to have thought in June that it was more valid, or you agreed more strongly with some of those statements than you had done in November, or whenever it was. Now does that have any kind of reality or is that just an arbitrary…
C: Sounds pretty arbitrary to me.
I: Right, OK. That's interesting to know because there were in fact…
C: The validity umm…
I: Yes, for instance, supposing I asked you to what extent you'd say the information provided by ALIS is a fair measure of performance, what would you say to that?
C: I guess it is. Yes, I think umm…Let me think. Our [name of syllabus] people for example, I think we did pretty well with them. We got very positive residuals for the first couple of years that we did it, and now we're just about hovering on the zero line, and I still feel it's a good course for those that do it. It's a good course, and it's become … it's made maths more popular. More people do it – and far more intermediate 'C's – and we've got a lot of students who are very much border-line in terms of pass or fail, and I feel, well, this is what has lowered our residual. But then it shouldn't do, because we're still comparing with everyone else who has gone in, who've come in with the same sort of grade. It's just that maybe in some schools there are a few people in a class that have those sorts of incoming GCSE averages whereas we end up sometimes with a whole group, more or less, that have got that sort of average.
I: So, if you have a lot in a particular group, you think there's a kind of group effect that …
C: Yes
I: Yes, it's very hard to know that, but yes. There are lots of things ALIS doesn't take account of, and that would be one of them of course.
C: I mean, in the last couple of years we've had upper sixth groups where really we've predicted that about half of them are not going to get it, and in most cases they've either given up or those half haven't got it. And you feel, you know, has it drawn down the mark for the others because we've spent so much time chasing them up, and the stress of classroom management? And not being able to push the better ones?
I: Right OK. So all sorts of things it doesn't pick up?
C: Because it's the ones with the middle GCSE range – or in the band 2 that you split them up into four groups don't you? So the ones in the bottom group often, some of them will do vastly better than expected, some worse, and then it's the one above and the next one up where I feel we're failing them a bit – they're not getting as good a residual as they should.
I: So do you look at individual students?
C: Mmm, oh yes. Yes I usually take the ALIS graph and plot their results as individuals on either side of the line and just see how it looks as individuals and as groups.

I: Well, I've got some other questions so let me just ask you these then. To what extent would you say that you see your students success or failure as being within your control?
C: Hmm. Half and half. There are so many others that have peer pressures and … Yes, I think you can … I also look at some other members of staff who don't have as much admin. as I do. I sometimes feel am I not doing as much preparation.
I: Right
C: And yet, you know, I see other staff who are sort of struggling with theirs and they are having a bit more time for theirs. It doesn't seem to create much difference in terms of the final result
I: That's interesting
C: Unless your results show otherwise [laughter].
I: Well, I don't know…Would you say that's something that has changed at all or is that constant, the amount of control you feel you have over it. Has your feeling about that changed?
C: It's lessened because of the size of the groups I think
I: Right OK
C: Last year I happened to have a group that went down in size to eleven and it was a dreadful group to begin with. It had a lot of poor students who weren't going to succeed so found it hard to concentrate and we actually managed to get rid of quite a few in the first year; by the second year it was a lovely group. And I feel – we got 4 'A's in the group – and I think that was because it came down to the size where they were all supportive and working well.
I: OK, now what about your own perception of your personal effectiveness as a teacher. How confident do you feel about that?
C: Umm. Is that different from the last question?
I: Yes, well, I mean you could feel that you were in control but not effective or that you were, I suppose if you were, if you think you were effective then that implies some measure of control but not necessarily. Well, maybe it isn't different then!
C: It very much depends on, you know, group by group again.
I: Right, so is it possible, I suppose what I'm getting at is it possible to say that some teachers are effective and some are not or is that just a gross over simplification?
C: Hmm. You know I've sort of struggled with this over many years because I see the different styles of all the people in the department you know from [name of teacher] through to whoever [laughs]. [Name of teacher] had a very different style to someone like [name of another teacher], for example. But quite often you'd look at the results and you wouldn't see a vast difference because I think [name of first teacher]'s students had to become very independent. So, it's hard to nail it down.
I: Yes, that's one of the things that I'm trying to look at by having evidence about particular individual teachers, to what extent that varies from year to year for instance? And it varies. It seems to be that although you'd expect for an individual teacher –because you've got less students than say a whole department of your sort of size – you'd expect it to vary more because smaller numbers tend to be more variable. But in fact it's about the same, I think. So I think the teacher effect is probably bigger than the department effect and that kind of cancels out the size effect. That's a provisional finding I think. But it still varies – well you know how much they vary from year to year. And of course the big question – I should be asking you questions really … The big question all the research on this tends to assume is that schools are responsible for it. You know, they talk about school effectiveness as if it was unproblematic and we know that it varies anyway, and how much is it under your control?
C: As I said the size of groups that certainly has made an effect. I mean chatting to someone at the [subject] meetings who teaches at [names school] he said the parents would be up in arms if I had to take more than twelve per group. So he has about 30 students doing it in about 4 groups or something.
I: Goodness. And what sort of size are your groups now?
C: Umm. 23 at the moment is the largest
I: This is starting size is it?
C: Yes.
I: Right
C: But we've got groups of twenty still in the upper sixth.

I: Oh right. So you hang on to most of them.

C: Not all of them. I mean some of the groups go down a bit but we've sort of jiggled around as to whether they choose Stats. or Mechanics. Yes, I've got about twenty doing Mechanics. So that's the other thing, I suppose when I sent you the info about different teachers because we re-arranged them in the second year. Oh, no, I think I did put that on didn't I?

I: Well, it was quite complicated yes

C: Yes, because they'd had sort of 0.25 of one teacher and …

I: Nobody else gave me that kind of detail, but I can use all that so that's good. One of the problems with it is where people do share a group, particularly if three people do it is quite hard to know what the effect of that is.

C: And you don't have the breakdown, like in the [name of syllabus], the Mechanics marks versus the Pure, because that's what our division is, and we have got those marks separately.

I: You have? Right, but I haven't. Yes, in theory, and probably if I'd thought of it early enough I could have got that.

C: But is that useful actually, because your Pure is poor it's not going to support your Mechanics either

I: Right, yes, so it's complicated.

C: And the Pure is often what they come in with, which is poor and they can sort of make up marks on the Applied because it's a slightly different type of assessment.

I: Mmm. It is complicated.

C: Yes it is

I: The more I think about it, the more complicated it seems. OK, I think that just about covers everything that I wanted to check.

C: Yes, I think I think different things, you know, which probably is the reason why I answered differently maybe the second time. I didn't look back on what I'd answered the first time, I don't think I'd even kept it.

I: No, no, of course not. I didn't expect you to. But you know if I'm using it – if I'm claiming that it measures anything – then I need to be able to justify the idea that it's kind of a … well, either that it's a reasonably robust thing or that there's some other reason for thinking it may have changed, I suppose. And it's all about your perception anyway so if you tell me you don't think it has changed then well, that's all useful evidence. [C laughs] Maybe not quite what I wanted to hear, but never mind

C: Perhaps I'll give you a new line of enquiry with the size of groups and things and well, the average GCSE level of the groups that sort of thing.

I: Yes, well those kinds of things have been looked at and the evidence is quite mixed. Some studies seem to find that what they call context effect – in other words the group being of a particular kind – that they do have an effect but then other studies have found that they don't.

C: Yes, it would be nice if you came up with conclusive evidence that A Level sets mustn't be larger than 16 or something

I: Yes well that's another thing. There's a whole complicated picture there, because teachers compensate, and they put people in groups – they put good students into bigger groups and that kind of thing – and they work harder with the bigger groups. So if you do well controlled studies where you allocate randomly and keep other things the same, then yes, there is quite a significant effect. But if you just take things as they are and look at correlations, then you find either there's no effect or sometimes the big classes do better. But, of course, we know …

C: Well, I can imagine having a big class actually of the [name of syllabus] we're teaching as the non-coursework one now. We're saying an 'A' or a 'B' because it's tougher algebra and it's just quite a different kettle of fish. I taught it last year for the first time and I just feel like it's going back to the old style where you can actually whip through things fairly quickly. Not with all of them – there are a few weaker ones – but in general …

I: So you dropped the [name of syllabus] then?

C: Yes, well [that syllabus] had not proved terribly successful ALIS-wise, which was one little bit of evidence that decided us to drop it. So we did one year of [that syllabus] modular but we weren't terribly happy with that, and they didn't have their own text

books and the best ones we saw were the [name of syllabus] books. So we were using those books but having to rewrite bits, leave bits out, and we thought well, this is a bit silly, lets change to [syllabus].

I: And have you had a set of results of those?

C: We've had one set of results yes which were 90/91%.

I: Is that good? Sounds good to me.

C: Yes. [Syllabus] this year went up to 87.5 whereas last year it had gone down to 75, the previous year they were about 87 and this is because of this large clump of sort of 'D', 'E', 'N' students. A lot of them fell off the bottom last year and a lot of them we just managed to get through this year.

I: Have you had anything from ALIS yet this year?

C: No.

I: Right. I don't think most people have.

C: No, I think we've opted to go for the mini ALIS without all the teaching and learning styles which we do internally anyway.

I: Ok, well thank you for that

C: Was it a help?

I: Yes, very much so. Thank you.

D (chosen for rating feedback as 'easy' to understand and large increase in rating of 'ALIS fairness')

D: … on the hoof, quite often they turn out to be the best lessons that you ever do, but there have been occasions when I've gone in and thought to myself no, no, that's not been very good at all and wish I could go back and do it again. I suppose that happens to everybody but there certainly is an element of that. It depends on, well take today for instance, I'm feeling quite good. I had two really good lessons this morning. Tomorrow, I'll probably have an extremely bad one and feel awfully depressed and give you a different answer. It's sort of patchy, but again it's a pressure thing. It does take time, especially in a subject like Physics, where you've got apparatus and that sort of stuff, it does take time to get yourself ready and tune into what the students want, especially in the early stages when they've all come from different backgrounds … But on the whole I do a good job. The students tell me I do a good job anyway. Perhaps it's me just being hypercritical of myself. Put it this way, I know I could do a better job – that's probably the best way of putting it. That's probably the key issue from my own point of view.

I: Now, about ALIS then. How fair would you say that the information it provides is, as a measure of performance?

D: Mmm. How fair…? The waters do tend to get muddied sometimes by, I mean what you can't build into it is the actual school they came from. Now we've had, possibly the worst student I've ever had in terms of understanding, I mean her brain was pre-Aristotelian basically, in terms of Physics. Actually, she had – well this was in the days when you didn't have 'A*'s – she actually had a double 'A' grade and she was appalling. Yet one of the most successful students I ever had, we actually took him on a bit dubiously on the basis of having a couple of C grades, and he ended up getting a grade 'A' A-Level. The original girl dropped out very quickly, she just couldn't hack it at all. So, of course the data we get is based on their GCSE result but that does tend to be coloured by where they come from. I mean some schools get very good grades, but basically what they turn out are unthinking robots. Other schools get very bad grades, but in point of fact what you've got is a very bright kid. So there is this sort of blurring of the edges so to speak. Now we can take the ALIS information and we can match it up against the schools and we can come up with our own assessment if you like of how well we're doing. But of course, that never comes through on the paper. You know when you get a computer analysis of this. So, one of the things we do actually, it's part of the routine here, we actually go through all our extreme cases. You know the ones who have done much better than expected and the ones who have done worse than expected and sit there and analyse them to try and see why. And we go into things like, we have things like their past school record and where they came from and it's a useful little exercise for us because it does, you know if you've

got a student from that school well, just watch the grade, it might not mean exactly what it says.

I: Right OK. And how about your general attitudes towards ALIS?

D: Well, yes. It's all quite positive. It was actually [name of colleague] and myself that went to [name of institution]. We both came back and wrote this paper saying 'wonderful stuff, lets get on board'. So, I've been a convert to it right from the word go. You have to be careful with it. It's like most … I mean statistics can be a very blunt weapon when you want it to be, it can be a very sharp one if you've been perceptive enough and you actually understand them properly. So, being a mathematician, it doesn't worry me but I know it worries people who aren't.

I: Is that your perception then, within the institution, that people in non-mathematical type subjects are perhaps less positive?

D: I think that if it gets introduced into an institution, that certainly was the case here. The first two or three years we were using it, it was very difficult to get the people who weren't mathematically trained to actually understand what the information was. That's got a lot better and it's part of the culture here, so no one has any real worries about it, but there are certainly certain aspects of it which people find a bit scary perhaps. They're not really quite sure what they are looking at and they'll come along and ask me – ask the scientists – what it actually means. It's fair enough.

I: Right OK. And would you say that your own personal attitude had changed at all since you've been using it?

D: What to ALIS?

I: Yes

D: Yes. Over the years the quality and the depth of the information we have been getting has improved remarkably. I mean there's a lot of it. In a sense it's another burden. In a sense you've got to sit and look at all of this stuff and analyse it and make some positive use of it but at least it's a useful thing to be able to do. I've certainly found it an interesting exercise. So, we tend to treat it fairly positively. It's a useful little tool as long as you are conscious of its limitations. I think that's the danger. If people don't understand the system terribly well it can be used as a blunt weapon, lacking finesse.

I: I wonder, can you remember at all about the feedback I sent you because it sounds as though you are quite an expert user of it and you do quite a lot of things that I'd either done for you or recommended that you do. So, did you pay much attention to that or was it really, 'Oh we've done this already'?

D: Oh, no. I can see the envelope actually. I can see where I've got your feedback. I must admit I did have a long hard look at it at the time but it was actually some time ago now. I need notice of that question basically! But it looked quite interesting. There was some interesting stuff in it. I can't quite remember what it was now.

I: Well, I think it sounds as though, I mean things like looking at individual students, picking out outliers. I did notice one of the comments you made about having information about students' performance in all their subjects. I did try to incorporate something on that, the average of all their residuals, because that again is something that I'd quite often wanted to know.

D: Yes. I mean one of things I, because we tend to do it in departments … It's a case of if I've got an extreme case student whose residual is minus 6 or something like that, his raw residual anyway, what I tend to do is make an effort to find out what other subjects he's doing, that's easy. Then go and see the other departments and see how he did in those. So, are we looking at someone who is just under performing in my subject or is he under performing in all of them? Because there is usually a different message there depending on the information that you get back. So, actually seeing what their average residual was overall their subjects and comparing that against your subject, it's an easy thing to do actually within an institution, it's not a difficult thing to actually do if you're just looking at extreme cases. But if you're actually trying to look at everybody in the group, now that would take an awful long time. So, if your computer can generate all of that then it's that much easier for us.

I: Right OK. Well, thank you very much for your time, you've been very helpful.

E (chosen for rating feedback as 'easy' to understand and increase in rating of 'ALIS fairness')

I: I'm quite interested to explore a bit more some of the things that you've said in the questionnaire, and to find out whether my interpretation of that is right or not. So, I've just got a few quick questions that I'd like to start off with. First of all, to what extent do you perceive your students' success or failure as being within your control?

E: That's a very, very difficult question. 'To some extent' is the answer to that I guess. I'm sorry.

I: That's OK. Fair enough.

E: I mean, sometimes we think we have a big input, or should I say I think I have a big input. On some occasions I think it's down to me largely and other times less so.

I: Right. When you say some occasions…

E: Well, sometimes I really think I've helped out students a lot and made a difference and other times I think that no matter what I'd done the student would have got an 'A' anyway, or would have failed anyway.

I: OK. So are you saying that depends on the student or it depends on…..

E: I think the teacher can make a big difference in some cases especially if the student is receptive to that. In other cases, the student's attitude makes it difficult for the teacher to make a big difference.

I: Right OK. And what about your personal, how confident do you feel about your personal effectiveness as a teacher?

E: Quite confident. Yes, quite confident. But without being arrogant about it I would hope.

I: OK. Sure. And would you say that had changed at all over say the last year or over a longer period?

E: Umm. Probably not. I mean gradually as one gets more experienced one gets more confident I guess. So only on a continuous level I guess but not in any huge step really.

I: OK. And what about ALIS then? How would you rate the fairness of the information provided by ALIS, fair as being a measure of performance?

E: I'm still not completely sure. Umm. I think it's very useful but I'm not completely convinced that the underlying correlation which exists can be fairly applied in each individual case. I mean, because I'm a Maths and Stats teacher, I'm used to using regression to analyse spreads of large amount of data, but I think it's a bit more difficult to apply it with any great amount of confidence to an individual student. And so we do get students who've performed to their up most level at GCSE because of the school they've gone to, or students whose GCSE results under reflect their potential. And so as a measure for judging added on value I think it works quite well but I think one's got to be careful of applying it in individual circumstances.

I: So, when you say useful then in what sense is it useful, beyond being fair?

E: Umm. I think if you're applying it to all the students, not to an individual but to all the sets of students that we've got then those things, those individualities probably even out. And so we can look and see where the positive residuals lie, in what ability ranges or in what sets even they lie and so on, if there are any common patterns and also patterns over time.

I: Right OK. And would you say that your view about that has changed over time, the fairness?

E: No, probably not. It's something I'm still willing to discuss with people about and haven't got firm opinions on.

I: Right OK. And could you sum up your general attitude towards ALIS? How positive would you say that was?

E: Mine personally is quite positive.

I: OK and has that changed at all?

E: I think it has probably become more positive and we're now talking about bringing in target minimum grades which would be based on ALIS and I'm hoping that that will be a good way forward as well.

I: Right OK. That's interesting because you said in one of the questionnaires that you were thinking about that.

E: Yes. It has now become college policy. We're going to introduce it as from this academic year.

I: OK. Well, great. I wonder if you have any general thoughts about the kinds of questions that were in the questionnaires? I don't know whether you can remember at all? And the kinds of interpretations that I might put on those? And whether, am I reading things into it or is it a fair reflection of what you thought or…..
E: I find it difficult to remember what all the questions were. I'm sorry.
I: OK fine. Right thank you very much I think that's been very helpful

F (chosen for rating feedback as 'impossible' to understand)

I: So, the first question then. I wonder to what extent you perceive your students' success or failure as being within your control?
F: Partly I would say. Do you want a percentage or…
I: Yes, if you could, that would be good.
F: I'd say about seventy five
I: Right OK, and would you say that that had changed at all say over the last year, or over any other longer time scale?
F: Yes, I would say, actually my seventy five is a bit high. I would say I had less control now than I did before.
I: Right, and why might that be then?
F: Largely because of their ability before they start.
I: Right so you have more control with more able students?
F: Yes.
I: OK. All right. So what about your feeling about your personal effectiveness as a teacher? How would you rate yourself there?
F: Umm. In what terms?
I: Well in general terms, I suppose, effectiveness … I'm probably most interested in how well your students do, I suppose.
F: Mmm. Acceptable [laughs]
I: Right. And would you say that had changed at all?
F: Yes. I would say that I don't feel as happy with the results as I used to.
I: Right. Why is that then?
F: Again, it's largely with the raw material, but also pressures of time and you know, not being able to do things as effectively as I used to.
I: Right. OK. Talking about ALIS then, to what extent would you say that the information they provide is a fair measure of performance?
F: Of the performance that they've done, it's fine. As an outcome, I mean as a prediction, I don't think it is.
I: Right OK.
F: Sorry about that.
I: No, it's OK. Do you want to say more about that?
F: Well, it just doesn't take personality into account, and it doesn't take, you know, time constraints, pressures that come on them during the two years they're here.
I: Right OK, yes. So, again your perception of the performance measurement, then, you say that's fine. Has that changed at all?
F: No, I don't think so.
I: Right OK. And what about … could you say what your general attitude to ALIS was?
F: I don't really take much notice of it, I have to say.
I: Right OK. You did comment – one of your comments on the questionnaire was something about how the stuff I sent you was impenetrable, or words to that effect.
F: I think that did have a bearing on it, yes.
I: I am quite interested in that because there were several other people that said that, and it did seem to me to make quite a difference to how … I mean the whole point of what I am doing is to try and see how people respond to it, either in terms of attitudes or in terms of their students. Some kind of knock-on effect on them. And obviously if it makes no sense at all to you, it seems that the effect would be different from somebody who maybe reads more into it or whatever. I don't know. So …
F: I think you're running into the difficulty of an artist faced with sheets and sheets of statistics. This is the main problem really.

I: So it's the numbers that are off putting?
F: It is really, yes.
I: OK. Is that something that you feel confronted with or is it something that you can avoid, do you think?
F: Well, my husband is a mathematician so I could have got him to go through it with me and explain it to me if I had been that interested, but I didn't. I don't know. I think if we had to do something with it, then I could probably manage it.
I: Right OK. So it's partly perhaps a question of seeing some value in it is it?
F: Yes, I think that's right, yes.
I: OK. I don't think there's anything else I need to ask you. That's very helpful.

Appendices for Chapter 7:

Project 2: Data Collection

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

13 February 1997

Dear Head of Department,

You have been selected, as part of a random sample of schools and colleges in the ALIS project, to receive the enclosed information about the students in your department. I hope you will find it interesting and useful.

If you have any comments on any of the following (or any other) issues I would be very happy to receive them:

- How useful did you find the extra information (ie the bits which are not part of the feedback you have already received from ALIS)?
- How useful did you find the predicted grades for 1997?
- Are the predicted grades broadly in line with your expectations?

Yours faithfully

Robert Coe

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

NOTES AND SUGGESTIONS FOR USING THE ENCLOSED PRINTOUTS:

1. 1996 RESULTS:

Notes for interpreting this information:

This printout contains information about the students who took A level in 1996.

The first section of it (from the word "results" in the first column) contains the following information about the students who achieved a grade in your subject and for whom an average GCSE score is available:

- their **surname**, **forename** and **sex** (columns 2 to 4)
- their **average GCSE points** at the beginning of the A level course (where A* = 8, A = 7, B = 6, C = 5, etc) (column 5)
- the **nearest** whole **GCSE grade** to this average (column 6)
- their **A level grade** (column 7)
- their (**st**andardise**d**) **residual** - a measure of how their grade compares with that of others who started with the same average GCSE score. This is therefore a measure of 'value added' performance (column 8)
- the **broad value added** category into which their residual score falls: "+" if it is better than average, "0" if it is broadly average, "-" if it is below average. These categories are designed so that in an 'average' department, roughly 25% should fall into each of the "+" and "-" categories, leaving 50% in "0".

The next section (from "averages" in column 1) contains average (mean) values for the group and for various subgroups. Note that:

- the average **A level grade** is coded on the UCAS scale: A = 10, B = 8, C = 6, D = 4, E = 2, N = 0, U = -2. If an average corresponds to an exact grade, that grade is printed; otherwise it may be interpreted by rounding to the nearest whole grade (eg anything between 5 and 7 counts as C).
- the symbol in the **broad value added** column indicates the significance of the average standardised residual in the previous column. 'Averages' from only one student are all coded "0"; averages coded "+" or "-" are sufficiently far (given the number of students involved) from the expected value, 0, that for a group of typical students, performing in line with all the others in the cohort, such a value would occur purely by chance one year in four. Averages coded "+ +" or "- -" would occur by chance one year in ten.
- the final column "**n**" indicates the number of students included in each average.

Averages are provided for the following groups of students:

- **All completing:** all the students listed under "results"
- **By avg GCSE:** all completing students whose average GCSE score on entry was closest to the "nearest GCSE grade" listed. For departments with a reasonable number of students, this gives an idea of whether value added performance varied with ability
- **By sex:** all completing students, separated into male and female
- **Stdnts who left:** this row gives the average GCSE score and the number of students who completed the ALIS questionnaire at the beginning of the course but for whom no A level grade is available. These students are then also listed by name in the final section of the

table (from "dropout" in column 1). In some institutions there are no such students and the last section and final average do not appear.

Suggestions for using this information:

Results:

Identify the students whose broad value added is coded "-". For each of them, ask:

- Is it fair to describe them as having underperformed?
- If so, can you account for their performance?
- Look at their **std residual** score. The number of grades by which they 'underperformed' (compared with other students who began with the same GCSE score) is between 1 and 1.5 times this value (eg a student with a score of -2.00 achieved between 2 and 3 grades below what might have been expected).

Consider these students collectively:

- Are there any common features among these students?
- What proportion of your students are in this category (compared with the expected quarter in an 'average' department)?

Repeat similarly for those coded "+".

Averages:

Into which broad value added category does the overall average ("All completing") fall?
Is this value consistent with:

- What you expected of the group before getting their exam results?
- How you felt when you did get their exam results?
- Your examination of the 'over' and 'under' performers as above?

How does the average **std residual** vary for students grouped by average GCSE? Are there any differences in performance between the most and least able? (be careful not to read too much into an average which is based on fewer than about 5 students - it is too sensitive to one or two extreme values to indicate a real trend reliably)

Does the average **std residual** differ for males and females? (again if you do not have 5 or more of each, any difference probably says more about the individuals in the small group rather than signifying a true 'gender effect')

How does the average GCSE score of those who left compare with that of those who completed the course?

Dropout:

Are the names listed a true reflection of those who started the course but did not complete it? If so, in each case why did they drop out?

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

NOTES AND SUGGESTIONS FOR USING THE ENCLOSED PRINTOUTS:

2. 1997 'PREDICTIONS':

Notes for interpreting this information:

This printout contains information about the students who will take A level in 1997.

All students who are currently in the ALIS database as entered for your subject in your institution are listed. (In some institutions some students may not yet have been entered, so the list may not be complete.). The table contains the following information:

- their **surname** and **forename** (columns 1 and 2).
- their **average GCSE points** at the beginning of the A level course (where A* = 8, A = 7, B = 6, C = 5, etc) (column 3).
- their **ALIS 'predicted' grade** on the UCAS scale (A = 10, B = 8, C = 6, D = 4, E = 2, N = 0, U = -2). This may be interpreted as the average grade achieved in that subject last year by students with the same average GCSE score. Although the value is given to two decimal places, it should be seen as only a very rough guide. On average, if you take the interval from a grade below to a grade above this value (ie + or - 2 either side of it) you would expect the actual grade achieved to fall in that range about 50% of the time (column 4).
- a suggested **min target grade**. This grade is the next whole grade above the 'predicted' value. It is therefore the minimum grade the student must achieve in order to contribute a positive residual to your department's average (assuming the relationship between average GCSE and A level grade is the same next year). Note that a very small number of students with extremely good GCSE grades (ie almost all A*s) will have a 'predicted' grade of higher than 10, and will thus have (small) negative residuals even if they achieve a grade A. These are coded ">A!" (column 5).

Suggestions for using this information:

Compare the **min target grade**s with your own predictions. You may be able to identify at this stage students who are in danger of contributing large negative residuals to your departmental average. You might like to share this information with the students. Bear in mind that:

- ALIS 'predictions' are very rough and contain a wide margin for error;
- if *all* your students achieve positive residuals, your departmental average residual will be extremely high. However, it's good to aim high!
- your average residual score from last year (see other sheet) could give an indication of the average you can expect this year. However, these too vary quite a lot, so unless you have figures for more than one year and/or a fairly large number of students (say 40 or more), you should not infer too much from it.

999 , Maths 1996 RESULTS Page 1

data type	surname	forename	sex	average GCSE points	nearest GCSE grade	A level grade	std residual	broad value added	n
results	XXXXX	AMANDA	F	5.67	B	U	-2.13	-	.
	XXXXXX	MARK	M	6.44	B	E	-1.52	-	.
	XXXXXXX	BEN	M	6.89	A	B	.09	0	.
	XXXXXX	IAN	M	5.89	B	D	-.30	0	.
	XXXXXXXX	KATHERINE	F	6.33	B	C	-.04	0	.
	XXXX	ALAN	M	6.56	A	A	1.10	+	.
	XXXXXXXXXX	PAUL	M	6.66	A	B	.31	0	.
	XXXXX	JENNIFER	F	6.44	B	B	.53	0	.
averages	All completing:		.	6.36	.	5.50	-.25	-	8
	By avg GCSE:		.	6.15	B	3.60	-.69	- -	5
	By avg GCSE:		.	6.70	A	8.67	.50	+	3
	By sex:	M	6.49	.	6.40	-.06	0	5	
	By sex:	F	6.15	.	D	-.55	-	3	
	Stdnts who left:		.	5.33	.	.	.	0	1
dropout	XXXXXXX	APRIL	F	5.33	C	.	.	.	.

INST_ID: 999 , English 1997 `PREDICTIONS' Page 1

surname	forename	average GCSE points	ALIS `predicted' grade	min target grade
XXXXXXXX	DEBRA	4.55	2.73	D
XXXXXXXXX	JENNIFER	6.40	7.20	B
XXXXXXX	MICHELLE	6.46	7.35	B
XXXXXX	EMMA	5.78	5.70	C
XXXXXXXX	STEVEN	6.11	6.51	B
XXXXX	DAVID	5.30	4.55	C
XXXXXXX	CLARE	5.56	5.17	C
XXXXXX	JENNIFER	4.50	2.63	D
XXXXXXX	KATHERINE	7.00	8.65	A
XXXXXXX	KATHERINE	5.56	5.17	C
XXXXXXXX	KELLY	6.00	6.24	B
XXXXXXX	NICOLA	6.44	7.31	B
XXXXX	TINA	6.44	7.31	B
XXXXXXX	JAYNE	6.44	7.31	B

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

13 February 1997

Dear Head of Department,

You have been selected, as part of a random sample of schools and colleges in the ALIS project, to be offered the opportunity to have your department's A level results for 1996 and 1997 analysed separately for each teaching group.

This analysis will be in addition to the full feedback that your institution gets from ALIS, and will be made available only to you, not to anyone else in your institution or elsewhere. It would also be possible additionally to produce a separate analysis for each teacher of the performance of all the students taught by them over the two years.

If you would like to receive this analysis then please complete the enclosed form and printout and return them to me at the above address. As this information is potentially quite sensitive, I think it is important that you discuss it with the members of your department and obtain the consent of all those involved. If the department is divided and some members want this analysis done and some do not, then it will be possible to provide the feedback only for those who wish to have it. It would also be possible to put the feedback for each teacher into a separate sealed envelope if desired.

Yours faithfully

Robert Coe

RESULTS 96

XXXXXXXXXXX COMPREHENSIVE SCHOOL
English
Group: 1
tchrs: AXX PXX

teaching group ID	surname	forename	average GCSE score	`predicted' A level score	A level grade	residual	std residual	broad value added
9999601	XXXXX	JONATHAN	6.60	7.62	10	2.38	1.07	+
	XXXXXX	SOPHIE	6.00	6.17	8	1.83	.82	+
	XXXXX	NICHOLAS	7.55	9.91	10	.09	.04	0
	XXXXXX	KATHRYN	7.27	9.25	10	.75	.34	0
	XXXXXXX	SIMON	6.40	7.14	10	2.86	1.29	+
	XXXX	LAURA	6.50	7.38	10	2.62	1.18	+
	XXXXXXXX	GARY	5.40	4.72	6	1.28	.58	0
	XXXXX	EMMA	5.67	5.36	8	2.64	1.19	+
	XXXX	HELEN	6.60	7.62	10	2.38	1.07	+
	XXXXXXX	DAVID	5.20	4.23	10	5.77	2.60	+
	XXXXXX	SCOTT	7.60	10.04	10	-.04	-.02	0
Mean			6.43	7.22	9	2.05	.92	
N	11		11					

TARGETS 97

XXXXXXXXXXX HIGH SCHOOL
French
Group: 1
tchrs: CX JX

teaching group ID	surname	forename	average GCSE score	`predicted' A level score	expected grade range	target minimum grade
99999701	XXXXXXX	LISAMARIE	5.00	1.65	N/E/D	E
	XXXXXXXX	JULIA	7.25	8.56	C-A	A
	XXXX	MADELEINE	6.00	4.72	E-C	C
	XXXXXXX	KATHRYN	7.20	8.40	C-A	A
	XXXXXX	HELEN	7.10	8.10	C-A	A
	XXXXXXX	MARIE	7.00	7.79	C-A	B
	XXXXX	NICOLA	6.60	6.56	D-B	B
	XXXXX	BRIDGET	5.80	4.11	E-C	C
	XXXXX	JO-ANNE	5.10	1.96	N/E/D	E

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

Dear colleague,

I enclose some information from the ALIS database about the A level class(es) you taught in 1996 and/or those who will take their A level in '97. The information has also been sent to anyone else who has shared the teaching of that class, but to no-one else.

There are two printouts for each 1996 exam class: a list of results and value added performance (headed 'RESULTS 96') and a scatterplot of those results (headed 'PLOT'). Each is accompanied by a sheet which explains what the information means and how you might use it.

If the 1997 exam entries in your institution were in the database at the time when I sent you the original lists, then you should also have a printout for each 1997 class, showing ALIS 'predicted' grades and targets (headed 'TARGETS 97'). Again, this is accompanied by a sheet of notes.

I have tried to assume no knowledge of ALIS and to explain what the figures mean in some detail; I hope you will feel that the resulting length and complexity are justified. At any rate, I would very much like to receive any feedback about the following (or any other) issues:

- Which parts of the information have you found useful?
- In what ways have you used it?
- In what ways did it add to what you have already had from ALIS?
- Was it over-complicated/incomprehensible/easily understood?

I hope to send you a similar analysis of the 1997 results as soon as they are available, so any comments received before then could be incorporated into the feedback and notes you get.

Yours faithfully

Robert Coe

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

NOTES AND SUGGESTIONS FOR USING THE ENCLOSED PRINTOUTS:

1. RESULTS 96:

Notes for interpreting this information:

This printout contains information about the students in a particular teaching group who took A level in 1996. It contains the following information:

- a unique identifier for the teaching group (column 1);
- their **surname** and **forename** (columns 2 and 3);
- their **average GCSE score** at the beginning of the A level course (where A* = 8, A = 7, B = 6, C = 5, etc) (column 4);
- their **'predicted' grade**. This is a point score which shows the average A level grade achieved by students in your subject in 1996 who started with the same average GCSE score. A level grades are coded on the UCAS scale: A = 10, B = 8, C = 6, D = 4, E = 2, N = 0, U = -2. Most values will not be exact grades but may be interpreted by rounding to the nearest whole grade (eg anything between 5 and 7 counts as C).(column 5);
- their actual **A level grade**, coded on the same scale (column 6);
- the **residual** score which is simply the difference between the actual and the 'predicted' grade. It is therefore a measure of how their grade compares with that of others who started with the same average GCSE score, ie an indication of 'value added' performance (column 7);
- their **st**andardise**d residual**. This is the result of dividing the residual by an appropriate factor so that values for different subjects and in different years can be fairly compared in terms of their frequency of occurrence. (Note for statisticians: the standardised residual is a Normal variable with mean 0, variance 1) (column 8);
- the **broad value added** category into which their residual score falls: "+" if it is better than average, "0" if it is broadly average, "-" if it is below average. These categories are designed so that in an 'average' department, roughly 25% should fall into each of the "+" and "-" categories, leaving 50% in "0" (column 9).

The row following the word 'Mean' gives the averages of the values in each column. The row following the letter 'N' gives the number of students listed and the number of (standardised) residuals calculated.

Suggestions for using this information:

Individual Results:

Identify the students whose broad value added is coded "-" (and similarly for those coded "+"). For each of them, ask:

- Is it fair to describe them as having underperformed (overperformed)?
- If so, can you account for their performance?
- Look at their **residual** score. If you halve this value (since one grade is two points on the UCAS scale) you get the number of grades by which they 'under(over)performed' (compared with other students who began with the same GCSE score).

Identify those with more extreme **residual**s, say above 4 or below -4. The performance of these students is more than two grades away from what might have been expected.

- In each case, can you account for their performance?

Consider these students collectively:

- Are there any common features among them?
- What proportion of your students are in each of the "+" and "-" categories (compared with the expected quarter in an 'average' group)?

Mean (average) values:

Look at the mean of the **residual**s. This value reflects the overall performance of the group. A value of more than 2 indicates that in this group performance was a over whole grade per student better than might have been expected. Is this value consistent with:

- What you expected of the group before getting their exam results?
- How you felt when you did get their exam results?
- Your examination of the 'over' and 'under' performers as above?

There may be some students whose results you feel you could legitimately exclude from the group average.

- On what basis would you justify excluding them?
- Recalculate the average without them. Does it make a difference?

(Slightly more complicated:)

Look at the mean of the **std residual**s and the number of values used to calculate it (N: this is the figure printed below the mean). These two values can be used to say how likely it is that an average as high (or low) as yours could be simply the result of a chance grouping of students (whose performance will naturally vary without having to attribute this to a 'teaching effect'). The critical values for a given number of students are shown below:

Number of students	5	6	7	8	9	10	12	14	16	18	20	25	30
Critical value (C)	0.88	0.80	0.74	0.69	0.65	0.62	0.57	0.52	0.49	0.46	0.44	0.39	0.36

The mean **std residual** of a random grouping of **N** students will be greater than this critical value (or less than minus it) in fewer than 5% of cases. Thus if the mean for your group is outside these limits, it is quite unlikely to be simply a chance event. However, what responsibility (if any) you as the teacher should take for this is very much open to argument.

On the other hand, if the mean for your group is less than the 'critical value' (ie between -C and C), then this is within the amount of variation expected purely by chance. Of course, the 5% level is an arbitrary choice: the larger the mean (for a given N), the less likely it is to be result simply of chance variation.

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

NOTES AND SUGGESTIONS FOR USING THE ENCLOSED PRINTOUTS:

2. PLOTS

Notes for interpreting this information:

This printout contains a scattergraph showing the 'predicted' and actual A level grades against their average GCSE score for the students listed on the 'RESULTS 96' sheet.

The positions of the '**A**'s on the plot represent the actual A level grades achieved: read down to find the average GCSE score on the horizontal axis (where A* = 8, A = 7, B = 6, C = 5, etc) and across to find the A level grade on the vertical axis (A = 10, B = 8, C = 6, D = 4, E = 2, N = 0, U = -2). Note that a single plotted '**A**' may represent more than one result.

The positions of the '**P**'s on the plot represent the 'predicted' A level grades, ie the average grade achieved by students with a given average GCSE score. Again, a single '**P**' may be more than one result.

The symbol '**$**' is plotted if an '**A**' and a '**P**' should appear in the same place.

Suggestions for using this information:

The '**P**'s should all lie on a straight line. Draw the best straight line you can through them.

- Are the '**A**'s mostly above or below the line? Note that the 'residual', listed in column 6 of the 'RESULTS 96' printout, is simply the distance above or below the line: '**A**'s above the line mean positive residuals, those below mean negative.
- Is there any difference in the pattern of '**A**'s as you move from left to right? For example, if the '**A**'s of those with lower GCSE scores are mostly above the line, while those with higher GCSE scores are below, then in that group the less able students seem to have done better - in terms of value added - than the more able.

You can identify patterns for different subgroups by colour-coding the '**A**'s, for example by sex, ethnic origin, previous school, etc. Identify the student(s) represented by each '**A**' from the 'RESULTS 96' list and colour the '**A**' appropriately.

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT

Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message); Fax: 0191 374 3506;
Email: r.j.coe@durham.ac.uk

NOTES AND SUGGESTIONS FOR USING THE ENCLOSED PRINTOUTS:

3. TARGETS 97:

Notes for interpreting this information:

This printout contains information about the students in a particular teaching group who will take A level in 1997. All students who are currently in the ALIS database as entered for your subject in your institution are listed. (In some institutions some students may not yet have been entered, so the list may not be complete. If you added any names to the original list I sent out, they will appear in this list only if I also know their average GCSE scores.). The table contains the following information:

- a unique identifier for the teaching group (column 1);
- their **surname** and **forename** (columns 2 and 3);
- their **average GCSE points** at the beginning of the A level course (where A* = 8, A = 7, B = 6, C = 5, etc) (column 4);
- their ALIS **'predicted' A level score** on the UCAS scale (A = 10, B = 8, C = 6, D = 4, E = 2, N = 0, U = -2). This may be interpreted as the average grade achieved in that subject last year by students with the same average GCSE score. Although the value is given to two decimal places, it should be seen as only a very rough guide (column 5);
- an **expected grade range**. This is an attempt to show the likely range around the 'prediction' into which the actual grade may fall. Even this range will capture only approximately half of the actual grades achieved (column 6);
- a suggested **target minimum grade**. This grade is the next whole grade above the 'predicted' value. It is therefore the minimum grade the student must achieve in order to contribute a positive residual to your group's average (assuming the relationship between average GCSE and A level grade is the same next year). Note that a very small number of students with extremely good GCSE grades (ie almost all A*s) will have a 'predicted' grade of higher than 10, and will thus have (small) negative residuals even if they achieve a grade A. These are coded "A+" (column 6);

Suggestions for using this information:

Compare the **target minimum grade**s with your own predictions. You may be able to identify at this stage students who are in danger of contributing large negative residuals to your group average.

You might like to share this information with the students, or even to use the **target minimum grade** to set targets for each student. Bear in mind that:

- ALIS 'predictions' are very rough and contain a wide margin for error, so it would be wrong to allow anyone's aspirations to be limited by these grades: if you think you can realistically set a higher target then do so;
- if *all* your students achieve positive residuals, your group's average residual will probably be very high. However, it's good to aim high!

DURHAM UNIVERSITY/ALIS FEEDBACK PROJECT
Durham University School of Education,
Leazes Road, Durham DH1 1TA
Tel: (0191) 374 3484 / 372 0168 (direct); 0191 374 3517 (message);
Fax: 0191 374 3506; Email: r.j.coe@durham.ac.uk

13 February 1997

Dear Head of Department,

You have been selected, as part of a ramdom sample of schools and colleges in the ALIS project to receive a free copy of the Target Setting and Monitoring Information System (TAMIS) disk and guide. I hope you will find it useful and would be interested to receive any comments you might have.

Yours faithfully

Robert Coe

Annex

Using Bootstrapping to get Confidence Intervals for Alpha

THE THEORY OF THE 'BOOTSTRAP'

The bootstrap is, in the words of Efron and Tibshirani (1993), 'a computer-based method for assigning measures of accuracy to statistical estimates' (p10). It was invented by Bradley Efron in 1979 (Efron, 1979) and uses repeated resampling from a sample to estimate standard errors and confidence intervals for any statistic that can be calculated from the sample, without having to make any assumptions about its sampling distribution. The following description of the bootstrapping process is based largely on Efron and Tibshirani (1993).

The bootstrap is of use when we have a sample of values which may be thought to have been drawn at random from a larger population, and a statistic which we can calculate for the sample and wish to estimate for the whole population. Let us call the sample $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ and the statistic $s(\boldsymbol{x})$. The statistic might be, for example, the mean (in which case we already have ways of estimating its standard error), or, as in the case here, Cronbach's Alpha (in which case we do not have any method for estimating the standard error). In fact, the bootstrap can be used to estimate the sampling variability of any statistic and hence provide an estimate of the accuracy of the corresponding population parameter. Let us denote the population parameter by S.

A *bootstrap sample*, $\boldsymbol{x}^*$, is obtained from $\boldsymbol{x}$ by randomly sampling, with replacement, from the original sample. Thus $\boldsymbol{x}^*$ also consists of n elements, all of them originally elements of $\boldsymbol{x}$, but each x_i from the original sample may have been chosen once, more than once, or not at all in the bootstrap sample. A large number (B) of bootstrap samples ($\boldsymbol{x}^*_1, \boldsymbol{x}^*_2, \ldots, \boldsymbol{x}^*_B$) are generated independently in a similar way. Corresponding to each bootstrap sample is a *bootstrap replication* of s, namely $s^*_j = s(\boldsymbol{x}^*_j)$, the value of the statistic, s, evaluated for the j^{th} bootstrap sample $\boldsymbol{x}^*_j$.

Bootstrap estimate of standard error of s

The bootstrap estimate of the standard error of the statistic, s, based on the sample, $\boldsymbol{x}$, is simply the standard deviation of the bootstrap replications ($s^*_1, \ldots, s^*_B$).

$$\text{se}^*(s) = \{ \sum_{j=1}^{B} [s^*_j - \text{m}^*(s)]^2/(B\text{-}1) \}^{½} \qquad \textbf{(1)}$$

Where m*(s) is the mean of the bootstrap replications.

A good estimate for this standard error can be achieved with values of B in the range 50 to 200, depending on the shape of the distribution of the s^*_js. Efron and Tibshirani (1993, p52) give a rule of thumb that even a small number of bootstrap replications, say, B=25, is usually informative, a good estimate of the standard error can often be obtained from B=50, and very seldom are more than B=200 replications needed. They also say that it is almost never a waste of time to display the histogram of the bootstrap replications.

Estimates of bias

If $s(\boldsymbol{x})$ is the sample mean and S the population mean, then we know that s is an unbiased estimator of S, regardless of the population distribution. That is to say that if enough random samples are drawn from the population and s calculated for each, the mean of the ss will approach S. However, for most statistics this is not the case and

there is some bias. The *bootstrap estimate of bias*, bias*(s), is defined to be the difference between the mean of the bootstrap replications, m*(s), and the original sample statistic, $s(\boldsymbol{x})$:

$$\text{bias}^*(s) = \text{m}^*(s) - s(\boldsymbol{x}) \qquad \textbf{(2)}$$

Once again, this is an estimate whose accuracy improves with the number of bootstrap replications, B. However, rather more replications are needed with this estimator than for the estimate of standard error. Efron and Tibshirani (1993) , p130) give an example where after B=400 the estimate of bias is still inadequate and a graph (p133) of the convergence of bias*(s) with increasing B shows that it can be erratic even with substantially larger values of B. Efron and Tibshirani (1993) give a formula for a 'better bootstrap bias estimate' which uses, instead of $s(\boldsymbol{x})$, a value based on the calculation analogous to that of the statistic s for a sample in which the x_i are weighted according to their frequency of occurrence in the bootstrap samples. This estimate is a significant improvement on bias*(s), but unfortunately its calculation is quite complex for a non-linear statistic such as alpha. Furthermore, the use to which an accurate estimate of bias can be put is somewhat problematic. Efron and Tibshirani (1993) warn of the dangers of using it to produce a bias-corrected estimate owing to the likelihood of the latter having substantially greater standard error.

Efron and Tibshirani (1993, p130) give a formula for a 95% confidence interval for the true bias in $s(\boldsymbol{x})$, i.e. for the limiting value of bias*(s) as $B \to \infty$:

$$[\text{bias}^*(s) - 2\text{se}^*(s)/\sqrt{B},\ \text{bias}^*(s) + 2\text{se}^*(s)/\sqrt{B}] \qquad \textbf{(3)}$$

This seems to provide a useful indication of whether the bias is likely to be large enough to need to worry about.

Confidence intervals for S

Having calculated the standard error of our statistic from the bootstrap samples, the simplest way to arrive at an estimated confidence interval for the population parameter is to use the appropriate *z*-value from a standard normal distribution. For example, a 95% confidence interval for the parameter estimated by *s* would be

$$[m^*(s)–1.96se^*(s), m^*(s)+1.96se^*(s)]. \quad \textbf{(4)}$$

This is described by Efron and Tibshirani (1993, p66) as the *standard confidence interval* for *S*. However, the accuracy of this interval depends on the sampling distribution of *s*, and its use is therefore reliant on asymptotic normal distribution theory. In fact, under most circumstances the parameter estimated by bootstrapping will have close to a normal distribution, particularly if the sample size, *n*, is large. A slight refinement of the standard confidence interval is to use the coefficient from a *t*-distribution instead of the normal distribution. This takes account of the uncertainty in the estimation of the standard error and thus widens the confidence interval slightly. If *n* is large, though, the percentiles of the *t*-distribution do not differ appreciably from those of the normal distribution.

Both these methods depend on the distribution of the bootstrap replications (s^*_1, …, s^*_B) being approximately normal. This can be a problem, however, depending on the particular statistic being estimated, especially if the original sample is small or the parent population particularly skewed. A more sophisticated estimate for the confidence interval is the *bootstrap-t interval*. This method requires an estimate for the standard error of $s(\boldsymbol{x}^*_j)$ for each of the bootstrap samples, and thus requires bootstrapping of each bootstrap sample if there is no simple standard error formula available. A modified version of this method is described by Efron and Tibshirani (1993) , p162) in which a smaller number of bootstrap samples are themselves bootstrapped to obtain a function relating the value of $s(\boldsymbol{x}^*_j)$ to its standard error. A transformation based on this function is then applied to a larger set of bootstrap replications so that their standard errors will be approximately equal, and these transformed values are used to estimate the *t* values required to give the appropriate proportions within the confidence interval.

An intuitively much simpler approach is to use the *percentile interval*. For this, the bootstrap replications are simply arranged in order and the values at the $[(100–p)/2]^{th}$ and $[100–(100–p)/2]^{th}$ percentiles taken as the lower and upper p% confidence limits. Thus, if a 95% confidence interval is required from a set of 1000 bootstrap replications, the 25^{th} and 975^{th} ordered values are the lower and upper limits. If $B.(100–p)/2$ and $B.100–(100–p)/2$ are not integers, Efron and Tibshirani (1993) suggest using the k^{th} largest and $[B+1–k]^{th}$ largest values of $s(\boldsymbol{x}^*_j)$, where k is the largest integer less than or equal to $(B+1)(100–p)/2$. The percentile interval estimate has a number of properties which make it a particularly good estimate with non-normal populations. First, unlike the standard confidence interval, it is not constrained to be symmetric – the main source of error in the latter. Secondly, it is *transformation-respecting*, in other words if the bootstrap replications $(s^*_1, \ldots, s^*_B)$ are transformed by some function and the confidence interval estimated using the transformed values, application of the inverse transformation to the interval limits will give the same result as would have been achieved by simply estimating the percentile interval directly. For a statistic such as the correlation coefficient, for which a transformation (Fisher's Z-transform) is known to make the distribution close to normal, the use of this transformation would make a considerable difference to both the 'standard confidence' and 'bootstrap-t' intervals. In fact, the use of such a transformation, where possible, to create a normal distribution makes the standard confidence interval quite accurate. Thirdly, the percentile interval is *range-preserving*, that is to say it cannot give values outside the allowable range for the statistic in question. This compares with the standard confidence interval which in the case of, say, a correlation coefficient, can give confidence limits above 1 or below -1.

Despite these advantages, however, the percentile interval can suffer from two kinds of inaccuracies. The first of these, bias, has already been mentioned. The statistic and sample may be such that the set of bootstrap replications are not 'centred' on the best estimate of the population parameter. The second, termed *acceleration* by Efron and Tibshirani (1993), refers to the amount by which the accuracy (i.e. standard error) of the sample estimate, s, varies with the true parameter value, S. If the variability or 'acceleration' is too great, then the distribution of the bootstrap replications is distorted.

An improvement on the simple percentile interval is referred to by Efron and Tibshirani (1993) as the *bias corrected and accelerated interval* estimate, or BC_a. This interval also takes its limits as elements of the ordered set of bootstrap replications, but adjusts the percentiles at which they are chosen to take account of bias and acceleration. For a p% confidence interval, the lower and upper limits are the $[100\alpha_1]^{th}$ and $[100\alpha_2]^{th}$ percentiles of $(s^*_1, \ldots, s^*_B)$ respectively, where:

$$\alpha_1 = \Phi\{b + [b+\Phi^{-1}(\alpha)]/[1-a(b+\Phi^{-1}(\alpha))]\}$$

$$\alpha_2 = \Phi\{b + [b+\Phi^{-1}(1-\alpha)]/[1-a(b+\Phi^{-1}(1-\alpha))]\}$$

(5)

Where $\alpha = (100-p)/200$, 'a' and 'b' are the acceleration and bias respectively, and Φ is the standard normal cumulative distribution function.

The bias, b, is defined as follows (note that this is not the same as bias*(s), defined above):

$$b = \Phi^{-1}\{\text{proportion of } s^*_j\text{s which are less than } s(\boldsymbol{x})\}$$

(6)

So if the median of the set of bootstrap replications is equal to s (the best estimate of S) then b will be 0. If the majority of the s^*_js are less than s, b is positive, if more, b is negative.

The acceleration, a, is somewhat more complex to calculate and less easily interpreted. Let $\boldsymbol{x}_{(i)}$ denote the n-1 members of the original sample which remain when the i^{th} element is deleted. Let J be the mean of the $s(\boldsymbol{x}_{(i)})$ (i.e. $\Sigma s(\boldsymbol{x}_{(i)})/n$) and $D_{(i)} = J - s(\boldsymbol{x}_{(i)})$ for each i. Then,

$$a = {}^1\!/_6 \,.\, \Sigma\,(D_{(i)})^3 \,.\, \{\Sigma\,(D_{(i)})^2\}^{-3/2}$$

(7)

Justification for this formula is given in Efron (1987).

The payoff for this extra complication is that the BC_a interval is often appreciably more accurate than the simple percentile interval. In terms of the size of the original sample, n, it can be shown that the error in coverage of the confidence interval derived by BC_a is inversely proportional to n, whereas the error for the standard and percentile methods decreases in proportion to $\sqrt{n}$ (Efron and Tibshirani (1993), p187). The BC_a method also retains the percentile method's advantages of being transformation-respecting and range-preserving, and of course is not constrained to give symmetric intervals.

Number of bootstrap samples required

It has already been noted that between 50 and 200 samples are generally sufficient for a satisfactory estimate of the standard error of s. It follows that estimated confidence intervals based on this standard error (i.e. the 'standard confidence interval') will not change appreciably as B increases. However, because the percentile methods are based on the bootstrap replications at the extremes of the distribution where there are fewer values, they are much more susceptible to sampling variations, and need a correspondingly larger set of replications on which to draw. Efron and Tibshirani (1993, p275) suggest that at least 1000 replications are often needed and generally work with B=2000.

OUTLINE OF THE BOOTSTRAPPING PROGRAM

Efron and Tibshirani (1993) provide details of software for bootstrapping that can be used as part of an integrated statistical environment such as S or S-PLUS. However, none of these programs were available in this research and it seemed to be both interesting and straightforward to write a custom made program. The program was written as a Visual Basic macro since the necessary functions were available in the spreadsheet Excel, and the data could easily be manipulated in an Excel file.

Setting up the file

Nine attitude scales had been constructed (see chapter 6) and it was hoped to be able to get an estimate of the variability of the value of Cronbach's alpha for each. Rather than recalculate all the formulae each time a new attitude scale was used, it was decided to create a generic program, and copy the relevant data into the file for each in turn.

Columns B to M of the worksheet contained the questionnaire response data, with column B containing the unique identifier for each questionnaire, columns C to L containing the responses for the items in the scale (with blanks where the scale contained fewer than 10 items), and column M contained the scale total (i.e. in each row, the sum of the values in columns C to L). Rows 1 and 2 were used for headings, so the data for the 72 questionnaires was in rows 3 to 74.

Obtaining bootstrap samples

Column A of the worksheet contained a random number function in each of the rows containing data:

```
=INT(72*RAND())+1
```

which calculated in each row a random integer between 1 and 72. For some of the attitude scales the data were incomplete, so rows had to be deleted and the '72' replaced by the appropriate number of cases. Thus, the contents of column A was a random sample of 72 numbers, each between 1 and 72, and if each number were used to select the questionnaire response corresponding to that unique identifier, it would be a bootstrap sample of questionnaires. The macro 'set_random' was used to paste this formula into the appropriate cells.

Each time the random numbers were recalculated a new bootstrap sample was therefore created. The numbers in column A were pasted into column N using 'paste special/values' in order to prevent them changing each time a calculation was done and thereby 'fix' the sample.

This pasting was carried out by the macro 'Zget_sample' (the prefix 'Z' was used for sub-routines which were only to be used when called up within another macro).

Columns O to Y were then filled using lookup functions to paste the data values in columns C to M corresponding to the unique identifier in column N. These functions were pasted in by the macro 'Xpaste_formulas' (the prefix 'X' was used to denote a macro that was used only once, in setting up the file, so that these would appear at the end of an alphabetical list of macros).

Calculation of alpha

Because the calculation of alpha involves a number of values for each member of the sample, the program is somewhat more complicated than it might be if it were to calculate a simpler statistic. The formula used to calculate alpha was:

$$\alpha = m/(m\text{-}1) \,.\, \{1 - \Sigma(\mathrm{Var}[u_i])\}/\mathrm{Var}[\Sigma(u_i)] \qquad \textbf{(8)}$$

Where $u_1, u_2, \ldots, u_m$ are the component items of a scale

Thus the program had to calculate the variance for each component item and for the scale derived by summing them. It also had to be able to work out how many items there were, since the lookup functions would have pasted zeros where the original questionnaire data was blank. Finally, it had to calculate the value of alpha for that bootstrap sample. These calculations were done by functions pasted in by the macro 'Xcalc_stats'.

Repeated bootstrap samples

Each time a value of alpha was calculated from a new bootstrap sample, it was pasted into the top of the next free column (column AA) and the whole of that column

moved down by one row. This was done by the macro 'Znext_alpha', which selected a new bootstrap sample.

The macro 'B_samples' repeatedly (B times) called up the macro 'Znext_alpha'. After running 'B_samples', column AA consisted of B bootstrap replications of alpha.

Estimating confidence intervals:

The bootstrap replications in column AA were sorted in order of size and the values at the 2.5% and 97.5% percentiles identified in order to estimate a 95% confidence interval. The mean, median, quartiles and standard deviation of the values were also found. The 'standard confidence interval' was calculated from the mean and standard deviation. Functions to calculate the acceleration and bias were pasted in anew each time a set of B bootstrap replications had been created, in order that their presence in the worksheet would not slow down the repetitions in running 'B_samples'. Using these values, the BC_a interval was found.

Running the program

The following steps were used to obtain a bootstrap sample:

1. Clear all the data and functions not needed for getting the bootstrap sample. Excel recalculates every function in the worksheet each time a new sample is created (hence generating new random numbers for selecting the sample), so it speeds things up appreciably to remove any that are not needed.
2. Copy the questionnaire responses for the relevant items into columns B to L. Delete any cases with missing data.
3. Recalculate the random number function (using the macro 'set_random') if the number of cases has changed.
4. Check that the correct number of random number functions have been pasted in and that no leftover functions remain from a previous run.
5. Close down any other applications running. Again, this makes the program faster.

6. Generate 1000 bootstrap replications of alpha and calculate the appropriate confidence intervals by running the macro 'B_samples'. (On a P100 processor this takes about 8 mins.)

Generating graphs

Following the advice of Efron and Tibshirani (1993, p53), and in line with general good practice, it was thought worth while to display the results of bootstrapping graphically, using two graphs. The first, a graph of cumulative frequency against alpha, was a simple way to represent the distribution of the set of bootstrap replications. A graph of the same relationship for the equivalent normal distribution (i.e. having the same mean and standard deviation) was overlaid for comparison. The 'raw' cumulative frequency graph automatically gives a smooth curve and from the comparison with the normal distribution it is possible to infer the shape of the distribution of the bootstrap replications. The second graph, a histogram of the distribution, makes this inference somewhat more obvious, although the variability in the heights of the bars can mask the overall pattern, even with as many as 1000 values. For this reason, the heights were 'smoothed' to produce a line which showed clearly the shape of the distribution. The smoothing was done by replacing each value with the median of itself and its two immediate neighbours and repeating this until none of the values changed further. Finally each value was replaced by the mean of itself and its two neighbours. In both types of replacement, end-points were left unchanged (see Tukey, 1977, for a full description and justification of the different ways of smoothing a bumpy distribution).

RESULTS

Confidence intervals for alpha for each construct

For each of the nine attitude constructs identified from the questionnaire responses (see Chapter 6), the value of Cronbach's alpha was calculated, based on all the complete responses to the items included in the scale (i.e. *n*, the number of cases).

The statistics for 1000 bootstrap replications for each of the constructs and the three kinds of confidence interval described above are shown in Table 48.

Table 48: Confidence intervals for each construct

SELF EFFICACY				
Original sample:		alpha: 0.779		n: 68
	B	mean	median	std error
Bootstrap samples:	1000	0.774	0.777	0.037
		lower:		upper:
Standard (95%) confidence interval:		0.702		0.846
Percentile interval:		0.694		0.836
Bias corrected accelerated interval:		0.712		0.848
	rank:	53		993
	bias:	0.1080	accel:	0.0419

ACHIEVEMENT ORIENTATION				
Original sample:		alpha: 0.789		n: 69
	B	mean	median	std error
Bootstrap samples:	1000	0.802	0.805	0.028
		lower:		upper:
Standard (95%) confidence interval:		0.746		0.857
Percentile interval:		0.741		0.848
Bias corrected accelerated interval:		0.680		0.827
	rank:	1		811
	bias:	-0.5476	accel:	0.0064

LOCUS OF CONTROL				
Original sample:	alpha:	0.832	n:	72
	B	mean	median	std error
Bootstrap samples:	1000	0.825	0.831	0.045
		lower:		upper:
Standard (95%) confidence interval:		0.737		0.914
Percentile interval:		0.721		0.896
Bias corrected accelerated interval:		0.731		0.902
	rank:	37		986
	bias:	0.0226	accel:	0.0386

ATTITUDE TO ALIS

Original sample:		alpha: 0.796		n: 65
	B	mean	median	std error
Bootstrap samples:	1000	0.770	0.774	0.042
		lower:		upper:
Standard (95%) confidence interval:		0.688		0.851
Percentile interval:		0.676		0.839
Bias corrected accelerated interval:		0.739		0.862
	rank:	225		1001
	bias:	0.5828	accel:	0.0207

FEEDBACK ANXIETY

Original sample:		alpha: 0.847		n: 72
	B	mean	median	std error
Bootstrap samples:	1000	0.846	0.849	0.031
		lower:		upper:
Standard (95%) confidence interval:		0.785		0.907
Percentile interval:		0.779		0.898
Bias corrected accelerated interval:		0.778		0.898
	rank:	23		975
	bias:	-0.0301	accel:	0.0086

SELF CONFIDENCE

Original sample:		alpha: 0.842		n: 69
	B	mean	median	std error
Bootstrap samples:	1000	0.832	0.837	0.046
		lower:		upper:
Standard (95%) confidence interval:		0.741		0.922
Percentile interval:		0.719		0.906
Bias corrected accelerated interval:		0.741		0.914
	rank:	41		986
	bias:	0.1434	accel:	-0.0179

FEEDBACK DESIRE				
Original sample:	alpha:	0.736	n:	72
	B	mean	median	std error
Bootstrap samples:	1000	0.726	0.735	0.064
		lower:		upper:
Standard (95%) confidence interval:		0.601		0.851
Percentile interval:		0.584		0.826
Bias corrected accelerated interval:		0.604		0.838
	rank:	40		988
	bias:	0.0050	accel:	0.0585

ALIS VALUE				
Original sample:	alpha:	0.792	n:	67
	B	mean	median	std error
Bootstrap samples:	1000	0.729	0.746	0.093
		lower:		upper:
Standard (95%) confidence interval:		0.548		0.911
Percentile interval:		0.527		0.876
Bias corrected accelerated interval:		0.652		0.913
	rank:	196		1000
	bias:	0.5948	accel:	-0.0427

ALIS FAIRNESS				
Original sample:	alpha:	0.689	n:	69
	B	mean	median	std error
Bootstrap samples:	1000	0.678	0.686	0.070
		lower:		upper:
Standard (95%) confidence interval:		0.540		0.815
Percentile interval:		0.519		0.790
Bias corrected accelerated interval:		0.524		0.798
	rank:	31		982
	bias:	0.0401	accel:	0.0067

It is apparent from considering the three confidence intervals for each estimate that there are not huge differences between them. The bias corrected accelerated interval may be the most accurate, but in most cases it is not much different from the other two. Particularly impressive is the standard interval which gives a fair approximation to the best estimate of the confidence interval, despite the asymmetry in the distribution, especially since this approximation could have been achieved with a much smaller number of replications. It seems that if a general idea of the size of

the confidence interval is required, rather than a precise estimate, the standard interval is likely to be adequate. On the other hand, if a precise estimate is needed, then it would probably be necessary to generate more than 1000 replications anyway.

Graphs

Cumulative frequency graphs and histograms for the distribution of bootstrap replications for each construct are shown in Figure 14 to Figure 31.

Figure 14: Cumulative frequency graph for Self Efficacy

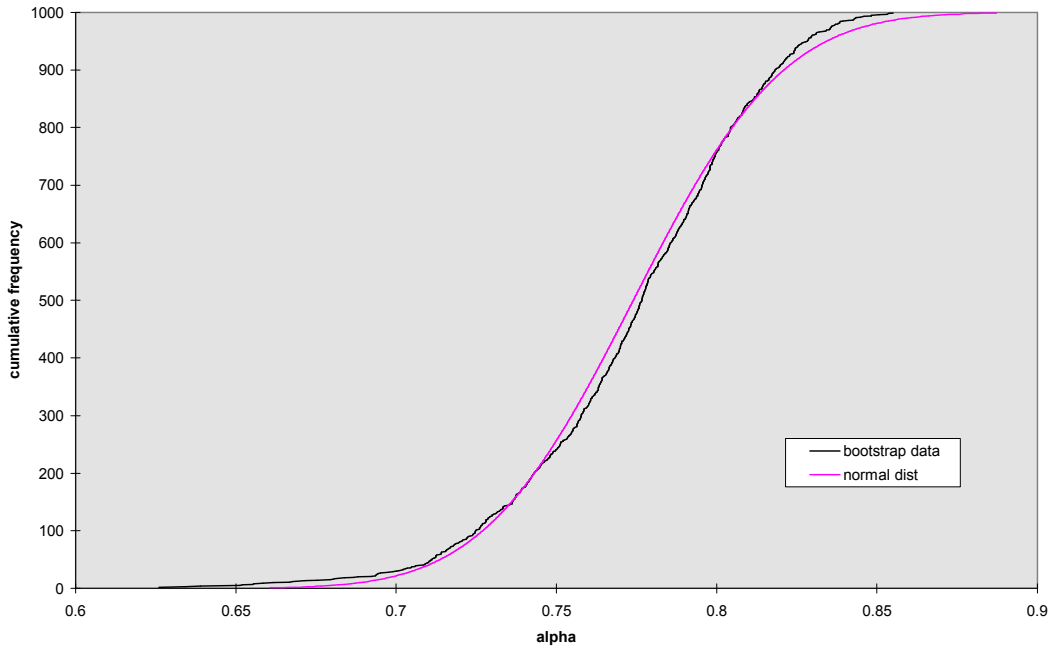


Figure 15: Histogram for Self Efficacy

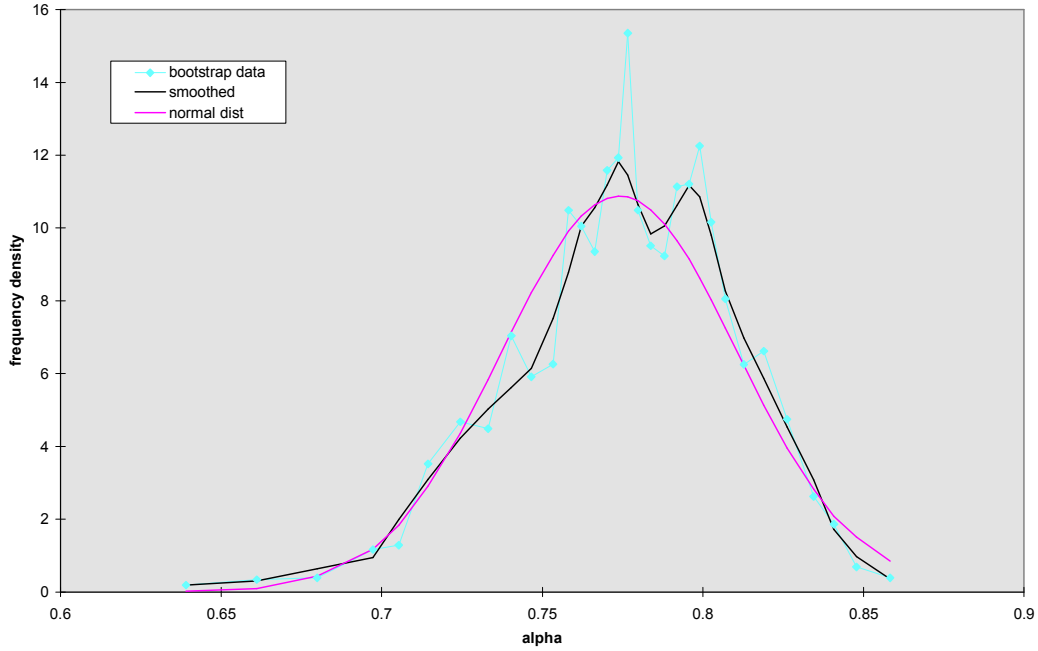


Figure 16: Cumulative frequency graph for Achievement Orientation

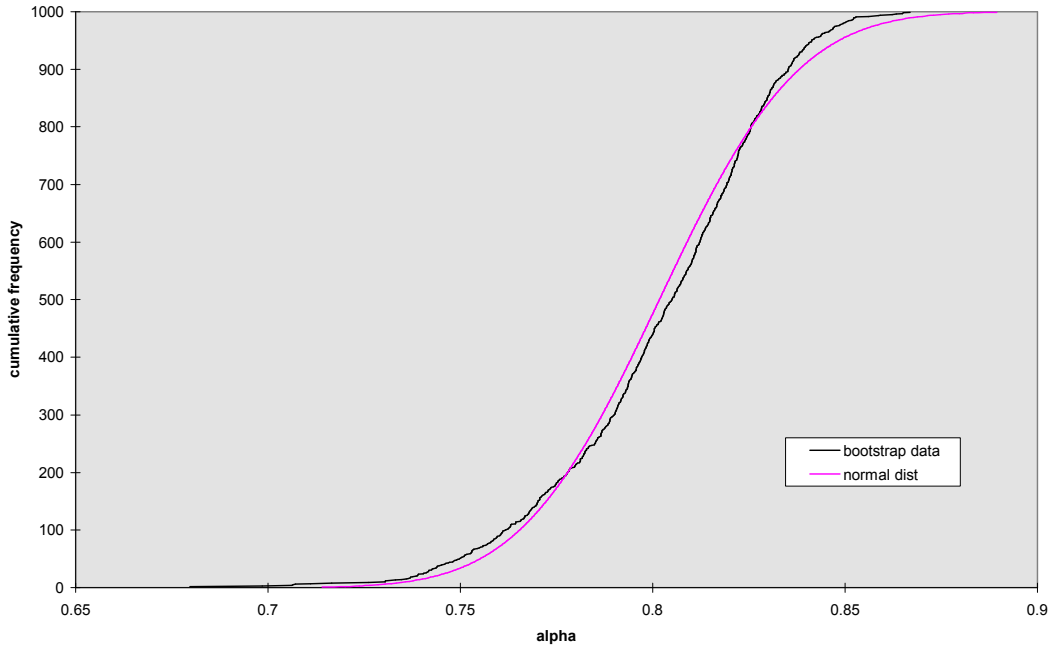


Figure 17: Histogram for Achievement Orientation

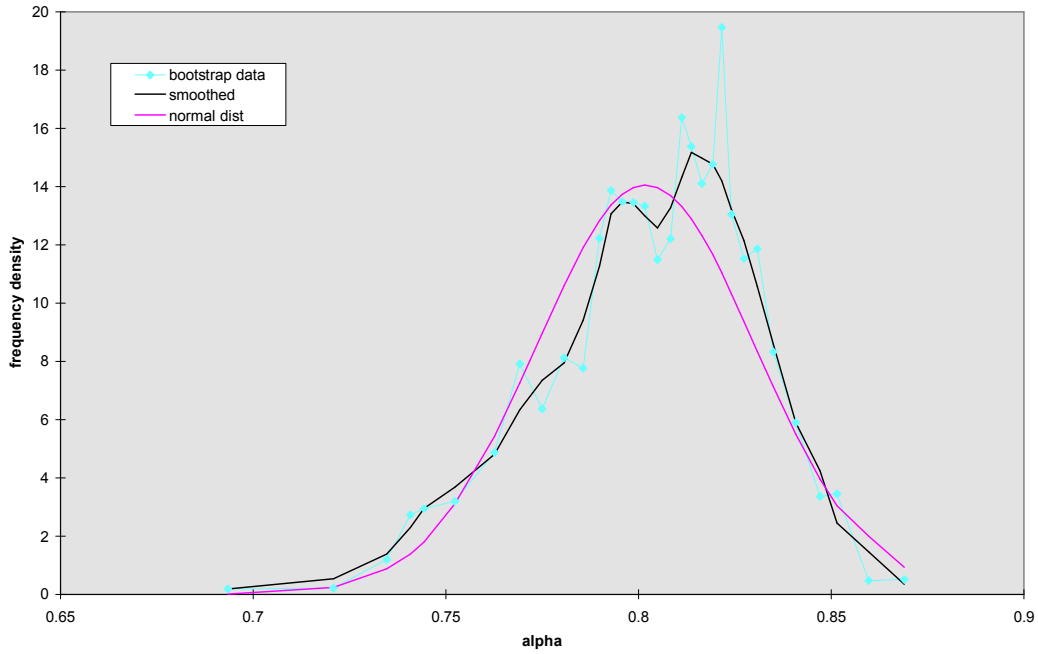


Figure 18: Cumulative frequency graph for Locus of Control

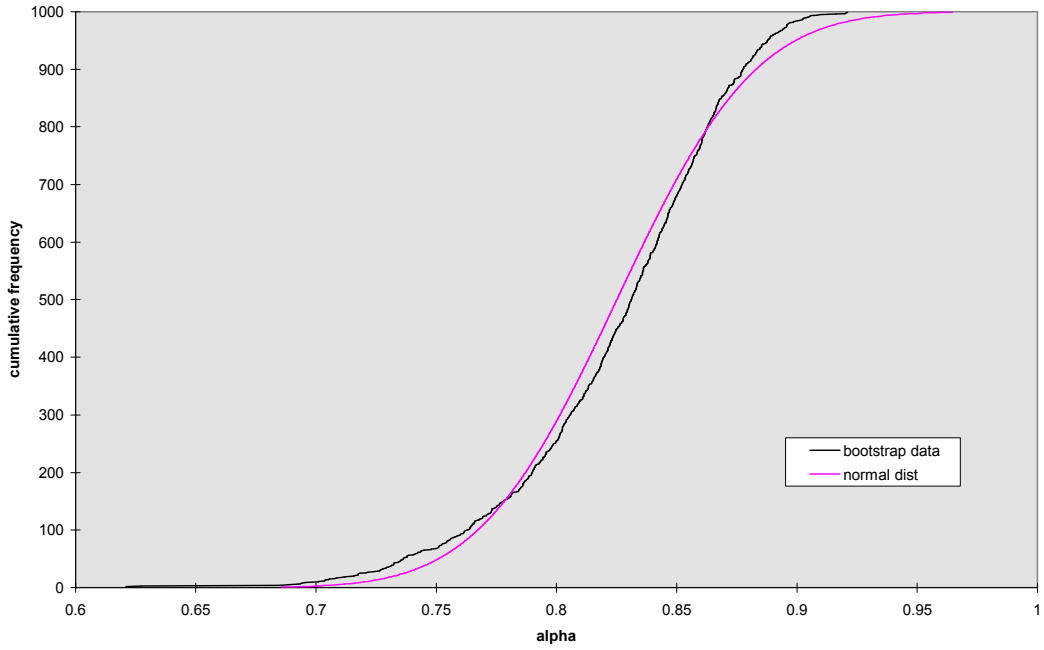


Figure 19: Histogram for Locus of Control

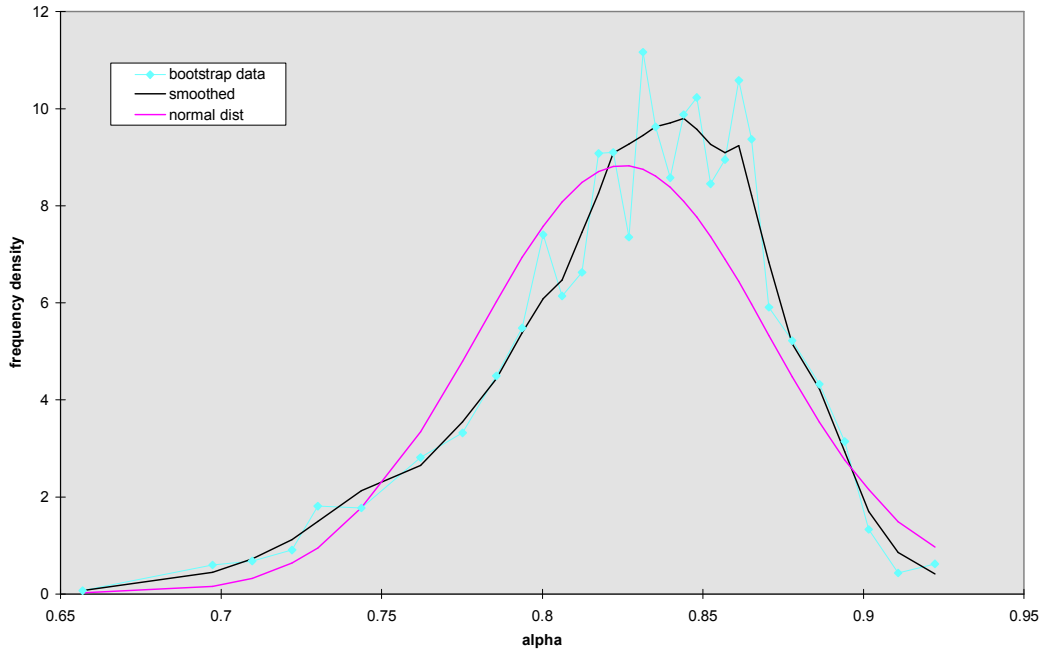


Figure 20: Cumulative frequency graph for Attitude to ALIS

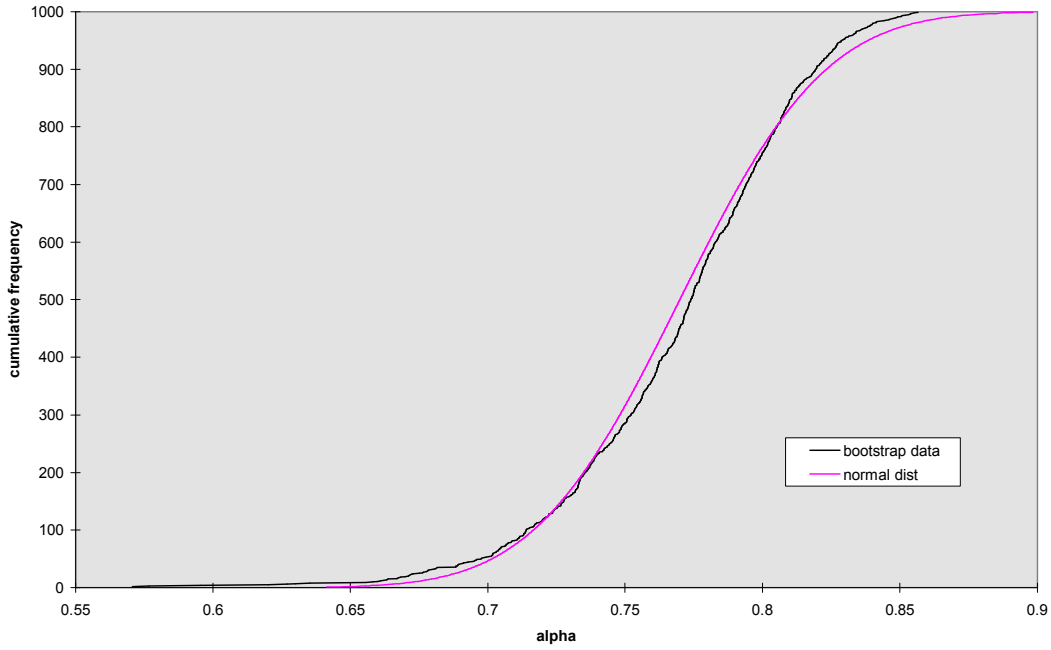


Figure 21: Histogram for Attitude to ALIS

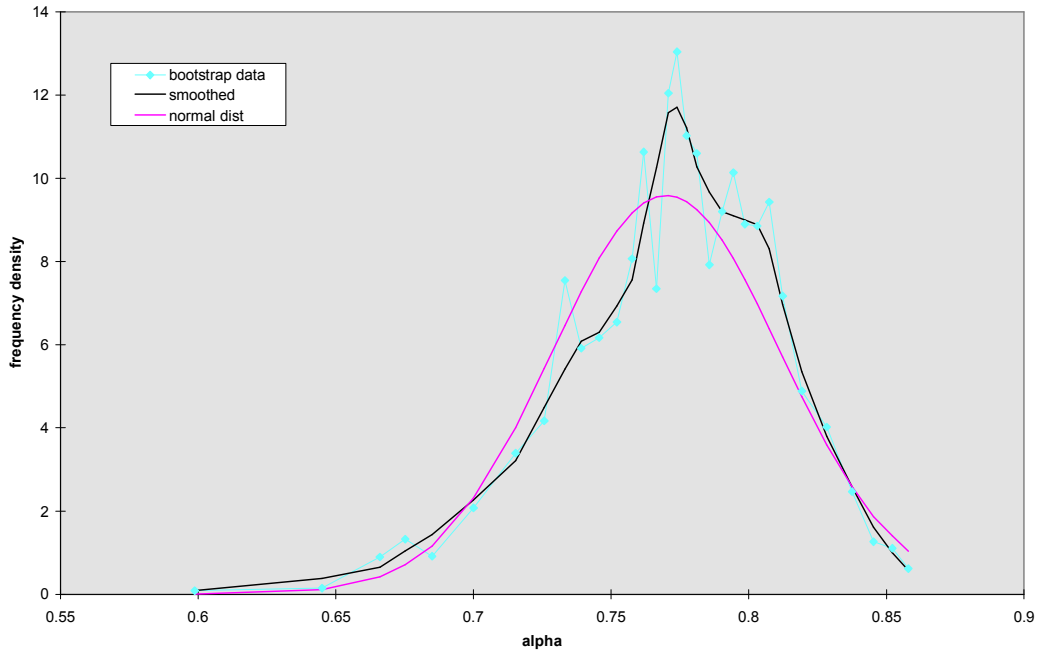


Figure 22: Cumulative frequency graph for Feedback Anxiety

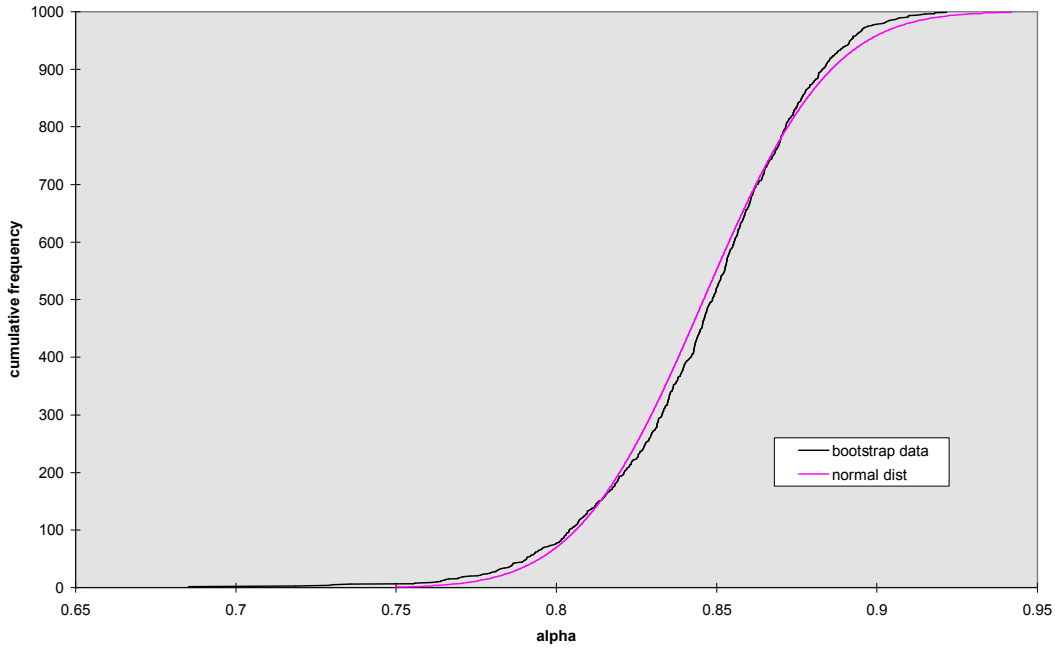


Figure 23: Histogram for Feedback Anxiety

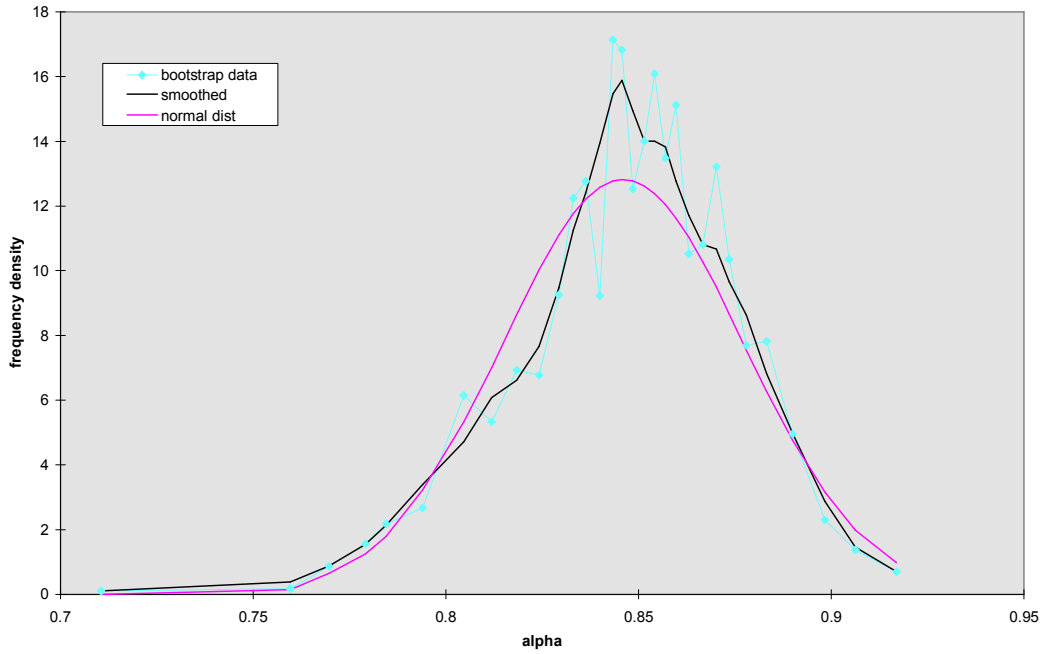


Figure 24: Cumulative frequency graph for Self Confidence

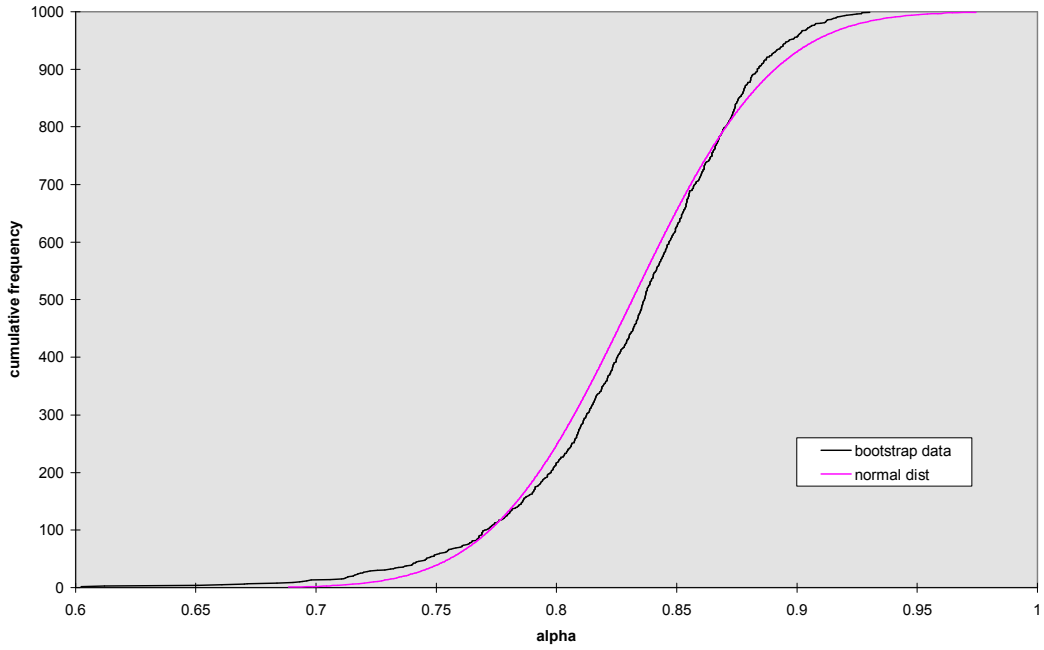


Figure 25: Histogram for Self Confidence

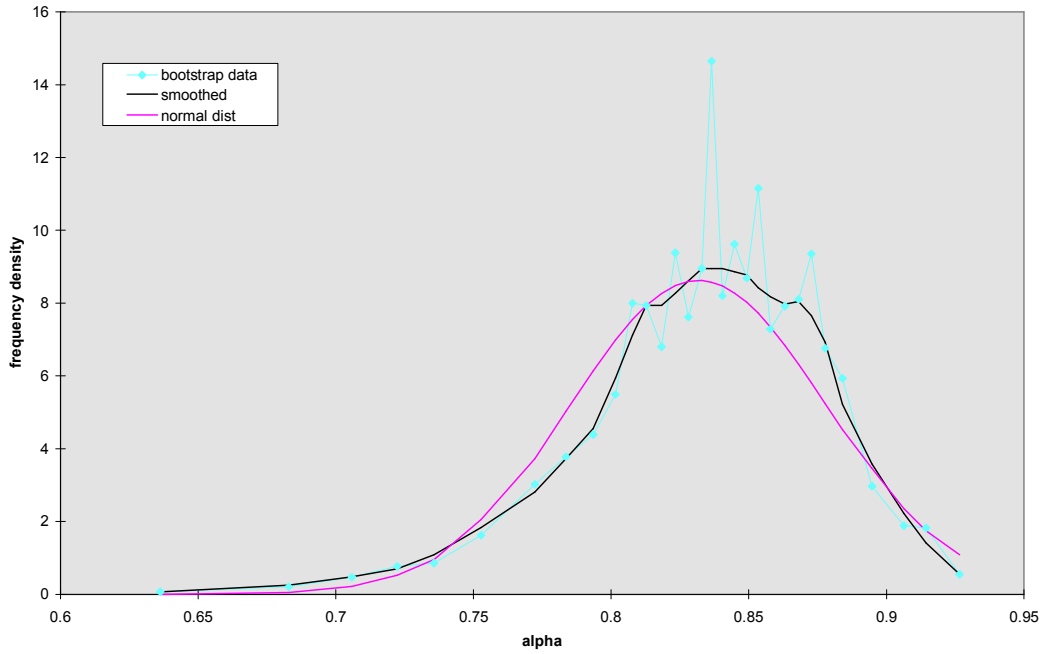


Figure 26: Cumulative frequency graph for Feedback Desire

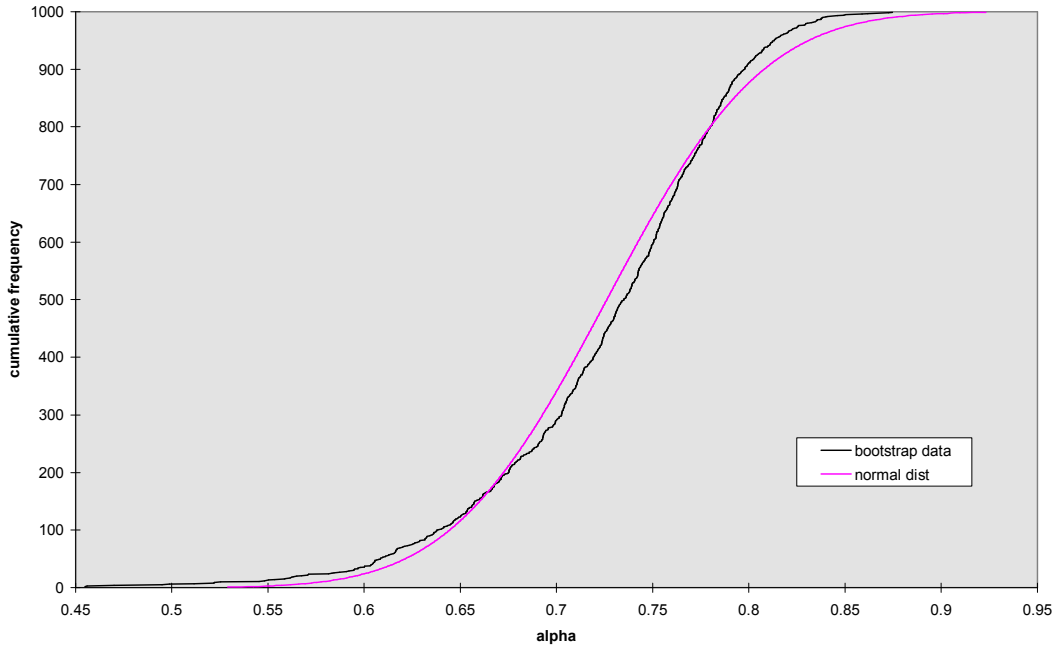


Figure 27: Histogram for Feedback Desire

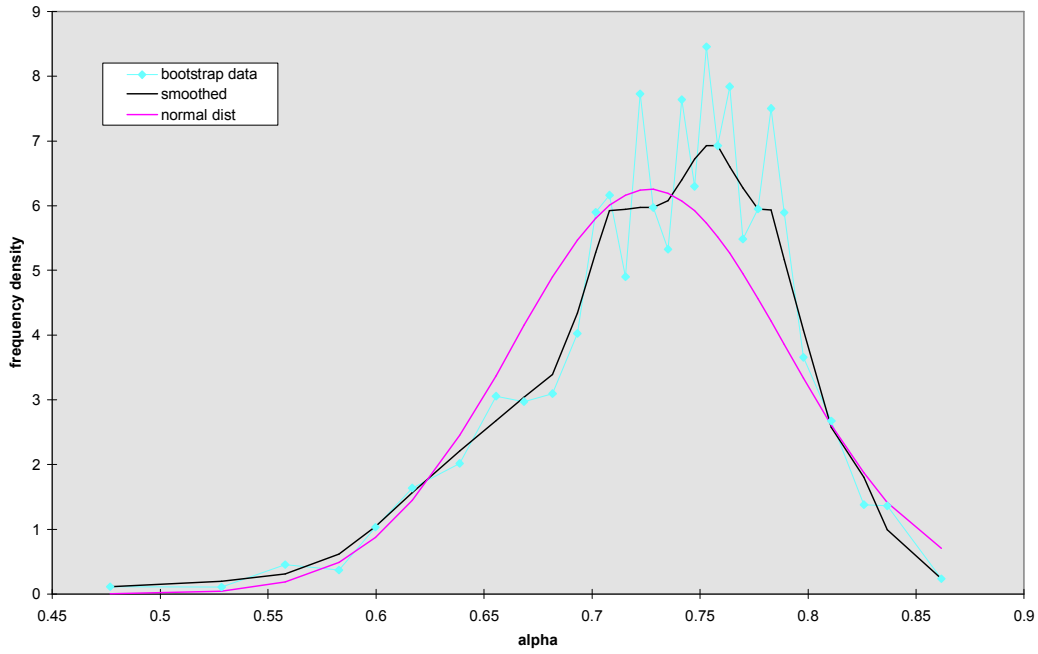


Figure 28: Cumulative frequency graph for ALIS Value

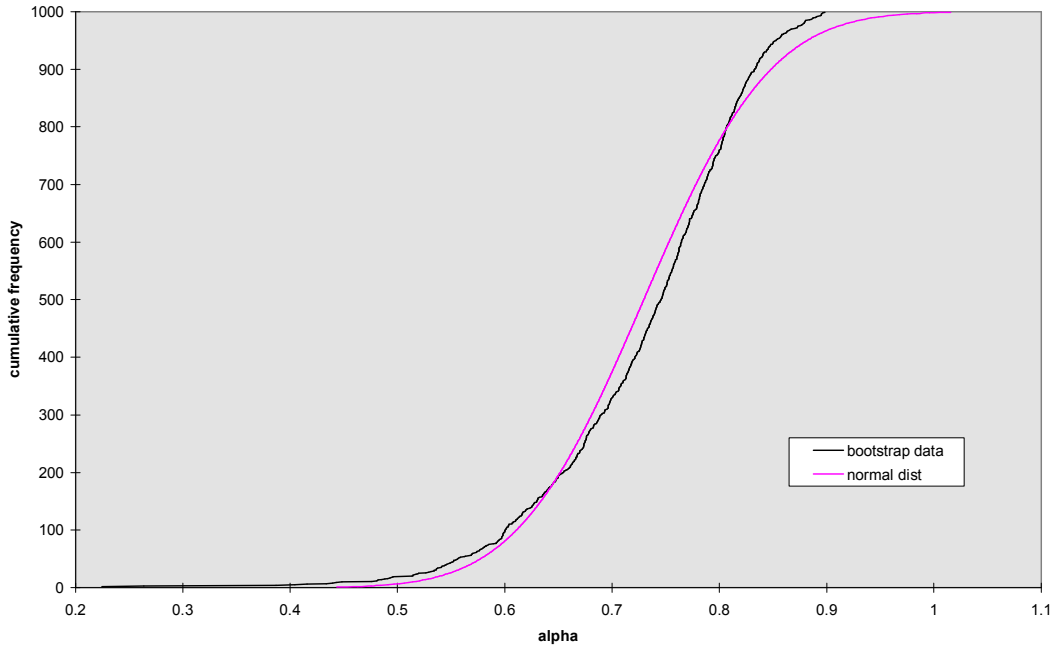


Figure 29: Histogram for ALIS Value

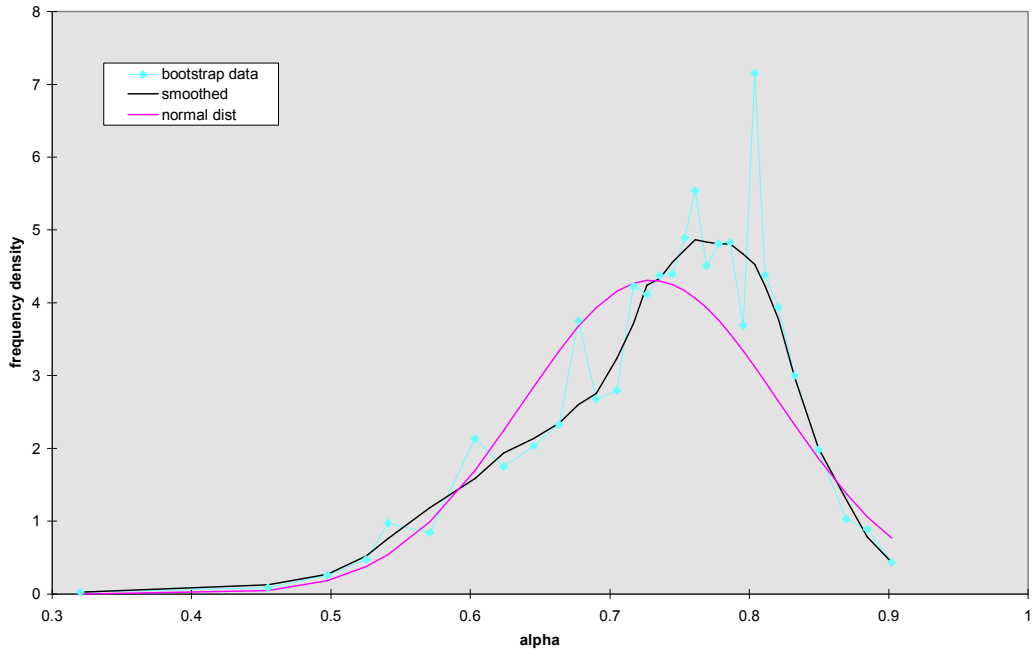


Figure 30: Cumulative frequency graph for ALIS Fairness

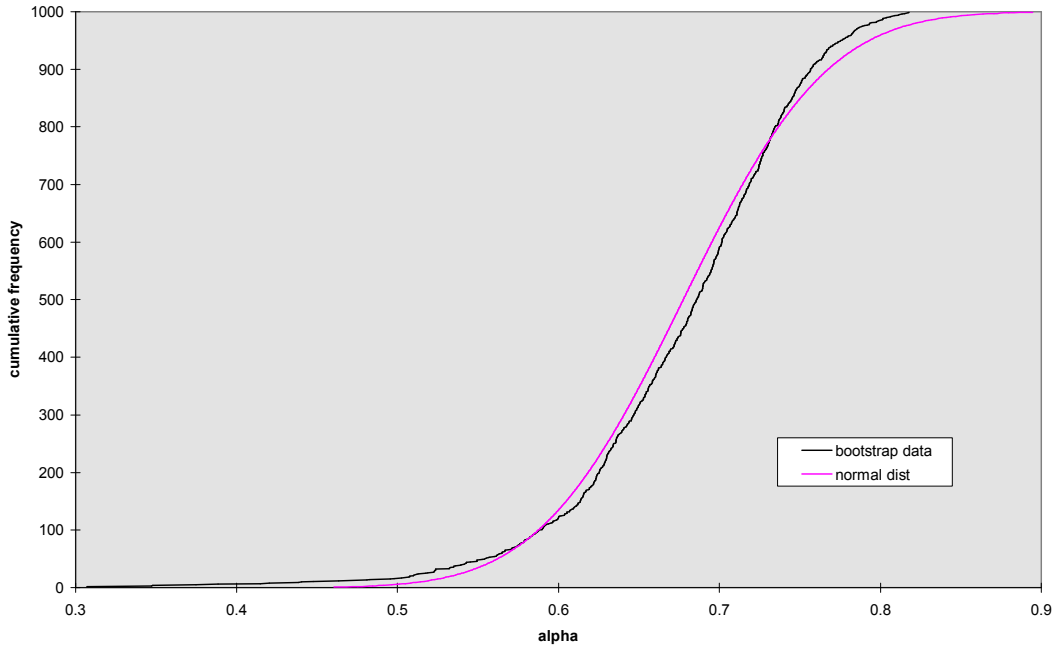
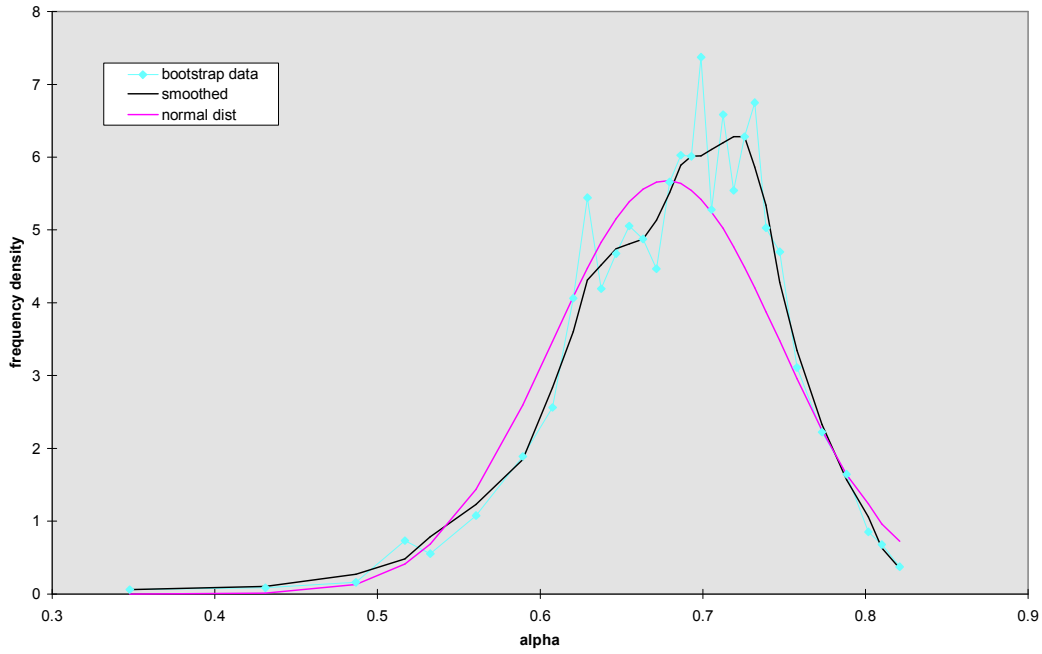


Figure 31: Histogram for ALIS Fairness



It can be seen from these graphs that all the constructs have generated much the same negatively skewed distribution.

References

Abramson, L.Y., Seligman, M.E.P. and Teasdale, J.D. (1978) 'Learned helplessness in humans: critique and reformulation'. *Journal of Abnormal Psychology*, 87, 1, 49-74.

Ammons, R.B. (1956) Effects of knowledge of performance: a survey and tentative theoretical formulation. *Journal of General Psychology*, 54, 279-299.

Anderson, C.A. and Jennings, D.L. (1980) 'When experiences of failure promote expectations of success: the impact of attributing failure to ineffective strategies'. *Journal of Personality*, 48, 3, 393-407.

Archer-Kath, J., Johnson, D.W., and Johnson, R.T. (1994) 'Individual versus group feedback in co-operative groups'. *Journal of Social Psychology*, 134 (5), 681-694.

Atkinson, J.W. and Feather, N.J. (Eds) (1966) *A Theory of Achievement Motivation*. New York: Wiley.

Azevedo, R. and Bernard, R.M. (1995) 'The effects of computer-presented feedback on learning from computer-based instruction: a meta-analysis.' Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1995.

Bandura, A. (1986) *Social Foundations of Thought and Action: a Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bangert-Drowns, R.L., Kulik, C.C., Kulik, J.A., and Morgan, M. (1991) The instructional effect of feedback in test-like events'. *Review of Educational Research*, 61, pp213-38.

Belson, W.A. (1981) *The Design and Understanding of Survey Questions*. Aldershot, UK: Gower.

Black, P. and Wiliam, D. (1998) 'Assessment and classroom learning', *Assessment in Education*, 5, 1,

Bloom, B.S. (1979) *Alterable Variables: The New Direction in Educational Research*. Edinburgh: Scottish Council for Research.

Boggiano, A.K. and Pittman, T.S. (1992) 'Divergent approaches to the study of motivation and achievement: the central role of extrinsic/intrinsic orientations', in A.K. Boggiano and T.S. Pittman (Ed.s) *Achievement and Motivation: a Social-Developmental Perspective*. U.K.: Cambridge University Press.

Bong, M. (1996) 'Problems in academic motivation research and advantages and disadvantages of their solutions'. *Contemporary Educational Psychology*, 21, 149-165.

Bosker, R. and Scheerens, J. (1989) 'Criterion-definition, effect size and stability: three fundamental questions in school effectiveness research', in B. Creemers, T. Peters and D. Reynolds (eds.) *School Effectiveness and School Improvement: Proceedings of the Second International Congress, Rotterdam, 1989*. Amsterdam: Swets and Zeitlinger.

Brandsma, H.P. and Edelenbos, P. (1992) 'School improvement through systematic feedback of pupil level data at the school and classroom level: an experiment.' Paper presented at the European Conference on Educational Research, Enschede, The Netherlands, June 1992.

Brandsma, H.P. and Edelenbos, P. (1998) 'The effects of training programmes for principals and teachers in secondary education: a quasi-experiment based on educational effectiveness indicators.' Paper presented at the International Congress of School Effectiveness and School Improvement, Manchester, January 1998.

Brinko, K. T. (1990) 'Optimal conditions for effective feedback' paper presented to American Educational Research Association annual conference, Boston MA, April 1990. EDRS No.: ED 326155.

Brinko, K. T. (1993) 'The practice of giving feedback to improve teaching: what is effective?'. *Journal of Higher Education*, 64, 5, pp574-93.

Brookover, W., Beady, C., Flood, P., Schweitzer, J. and Wisenbaker, J. (1979) *School Social Systems and Student Achievement: Schools Can Make a Difference*. New York: Praeger.

Butler, R, (1988) 'Enhancing and undermining intrinsic motivations: the effects of task-involving and ego-involving evaluation on interest and performance'. *British Journal of Educational Psychology*, 58, 1-14

Cameron, J., and Pierce, W.D. (1994) 'Reinforcement, Reward and Intrinsic Motivation: a meta-analysis'. *Review of Educational Research*, 64, 3, 363-423.

Camilli, G. and Hopkins, K.D. (1978) 'Applicability of chi-square to 2x2 contingency tables with small expected frequencies', *Psychological Bulletin*, 85, 163-167.

Camilli, G. and Hopkins, K.D. (1979) 'Testing for association in 2x2 contingency tables with very small sample sizes', *Psychological Bulletin*, 86, 1011-14.

Carroll, J.B. (1963) 'A model of school learning'. *Teachers College Record*, 64 (8) 723-33.

Carver, R. (1978) 'The case against statistical significance testing'. *Harvard Educational Review*, 48, 378-399.

Cliff (1993) 'Dominance Statistics – ordinal analyses to answer ordinal questions' *Psychological Bulletin*, 114, 3. 494-509.

Cohen, J. (1969) *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

Cohen, J. (1990) 'Things I Have Learned (So Far)'. *American Psychologist*, 45, 12, 1304-1312.

Cohen, J. (1994) 'The Earth is Round ($p<.05$)'. *American Psychologist*, 49, 12, 997-1003.

Cohen, L., and Mannion, L., (1994) *Research Methods in Education* (4th edition). London: Routledge.

Cohen, P.A. (1980) 'Effectiveness of student-rating feedback for improving college instruction: a meta-analysis of findings'. *Research in Higher Education*, 13, 4, 321-341.

Coleman, J.S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. and York, R. (1966) *Equality of Educational Opportunity*. Washington DC: US Government Printing Office.

Costanzo, P.R., Woody, E., and Slater, P. (1992) 'On being psyched up but not psyched out: an optimal pressure model of achievement motivation', in A.K. Boggiano and T.S. Pittman (Ed.s) *Achievement and Motivation: a Social-Developmental Perspective*. U.K.: Cambridge University Press.

Creemers, B. (1994) *The Effective Classroom*. London: Cassell.

Creemers, B. (1996) 'The school effectiveness knowledge base' in D. Reynolds, R. Bollen, B. Creemers, D. Hopkins, L. Stoll and N. Lagerweij (eds.) *Making Good Schools: Linking School Effectiveness and School Improvement*. London: Routledge.

Cronbach, L.J. and Furby (1970) 'How should we measure "change" – or should we?' *Psychological Bulletin*, 74, 68-80.

Cuban. L. (1984) 'Transforming the frog into a prince: effective schools research, policy and practice at the district level'. *Harvard Educational Review*, 54 (2) 129-151.

DeCharms, R. (1968) *Personal Causation*. New York: Academic Press.

Deci, E. L., and Ryan, R.M. (1985*) Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum.

Deci, E. L., and Ryan, R.M. (1987) 'The support of autonomy and the control of behavior'. *Journal of Personality and Social Psychology*, 53, 6, 1020-37.

Deci, E. L., and Ryan, R.M. (1992) 'The initiation and regulation of intrinsically motivated learning and achievement', in A.K. Boggiano and T.S. Pittman (Ed.s) *Achievement and Motivation: a Social-Developmental Perspective*. U.K.: Cambridge University Press.

Deci, E.L., Vallerand, R.J., Pelletier, L.G., and Ryan, R.M. (1991) 'Motivation and Education: the self-determination perspective'. *Educational Psychologist*, 26, (3&4), 325-346.

DfEE (Department for Education and the Environment) (1997) *Excellence in Schools*. London: HMSO.

Edmonds, R. (1979) 'Effective schools for the urban poor'. *Educational Leadership*, 37 (1) 15-27.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. NY: Chapman and Hall.

Fitz-Gibbon C.T. and Vincent, L. (1994) *Candidates' Performance in Public Examinations in Mathematics and Science*. London: SCAA.

Fitz-Gibbon C.T. and Vincent, L. (1997) 'Difficulties regarding subject difficulties: developing reasonable explanations for observable data', *Oxford Review of Education*, 23, 3, 291-298.

Fitz-Gibbon, C.T. (1985) 'A level results in comprehensive schools: the COMBSE project, year 1'. *Oxford Review of Education*, 11, 1, 43-58.

Fitz-Gibbon, C.T. (1992) 'School effects at A level: Genesis of an information system?', in D. Reynolds and P. Cuttance, *School Effectiveness: Research, Policy and Practice*. London: Cassell.

Fitz-Gibbon, C.T. (1996) *Monitoring Education: Indicators, Quality and Effectiveness*. London: Cassell.

Fitz-Gibbon, C.T. (1997) *The Value Added National Project: Feasibility Studies for a National System of Value Added Indicators (Final Report)*. London: SCAA.

Fitz-Gibbon, C.T., Tymms, P.B. and Hazelwood, R.D. (1989) 'Performance indicators and information systems' in D. Reynolds, B. Creemers, and T. Peters (1989) S*chool Effectiveness and Improvement*, proceedings of the First International Congress of School Effectiveness and Improvement, London, 1988. School of Education, University College of Wales, Cardiff.

Forsterling, F. (1985) 'Attributional retraining: a review'. *Psychological Bulletin*, 98, 3, 495-512.

Fortier, M.S., Vallerand, R.J., and Guay, F. (1995) 'Academic motivation and school performance: towards a structural model'. *Contemporary Educational Psychology*, 20, 3, 257-274.

Fortier, M.S., Vallerand, R.J., and Guay, F. (1995) 'Academic motivation and school performance: towards a structural model'. *Contemporary Educational Psychology*, 20 (3) 257-274.

Frey J.H. (1983) *Survey Research by Telephone* (2nd edition.). Newbury Park, CA: Sage.

Galloway, D. and Rogers, C. (1994) 'Motivational style: a link in the relationship between school effectiveness and children's behaviour?', *Educational and Child Psychology*, 11, 2, 16-25.

Galloway, D., Leo, E.L., Rogers, C., and Armstrong, D. (1996). 'Maladaptive motivational style: the role of domain specific task demand in English and mathematics'. *British Journal of Educational Psychology*, 66, 197-207.

Glass, G.V., McGaw, B. and Smith, M.L. (1981) *Meta-Analysis in Social Research*. London: Sage.

Goldstein, H. and Cresswell, M. (1996) 'The comparability of different subjects in public examinations: a theoretical and practical critique' *Oxford Review of Education*, 22, 4, 435-442.

Good, T.L. and Brophy, J.E. (1986) 'School Effects' in M.C. Wittrock (ed.) *Handbook of Research on Teaching* (3rd edn), pp570-604. New York: Macmillan.

Gray, J., Jesson, D. and Sime, N. (1990) 'Estimating differences in the examination performance of secondary schools in six LEAs: a multilevel approach to school effectiveness' *Oxford Review of Education*, 16, 2, 137-158.

Gray, J. (1995) 'The quality of schooling: frameworks for judgement' in J. Gray and B. Wilcox, *Good School, Bad School: Evaluating performance and encouraging improvement*. Buckingham: Open University Press.

Gray, J., Jesson, D. and Jones, B. (1986) 'The search for a fairer way of comparing schools' examination results', *Research Papers in Education*, 1, 2, 91-122.

Gray, J., Jesson, D. and Reynolds, D. (1996) 'The challenges of school improvement: Preparing for the long haul', in J. Gray, D. Reynolds, C. Fitz-Gibbon and D.

Jesson, *Merging Traditions: The Future of Research on School Effectiveness and School Improvement*. London: Cassell.

Gray, J., Jesson, D., Goldstein, H., Hedger, K. and Rasbash, J. (1995) 'The statistics of school improvement: establishing the agenda' in J. Gray and B. Wilcox, *Good School, Bad School: Evaluating Performance and Encouraging Improvement*. Buckingham: Open University Press.

Gray, J., Reynolds, D., Fitz-Gibbon, C. and Jesson, D. (1996) *Merging Traditions: The Future of Research on School Effectiveness and School Improvement*. London: Cassell.

Greenwald, A.G., Pratkanis, A.R., Leippe, M.R., and Baumgardner, M.H. (1986) 'Under what conditions does theory obstruct research progress?'. *Psychological Review*, 93 (2), 216-229.

Hall, G.E. and Loucks, S.F. (1977) 'A developmental model for determining whether the treatment is actually implemented', *American Educational Research Journal*, 14, 3, 263-276.

Harackiewicz, J.M., Abrahams, S. and Wageman, R. (1987) 'Performance evaluation and intrinsic motivation: the effects of evaluative focus, rewards and achievement orientation'. *Journal of Personality and Social Psychology*, 53, 6, 1015-23.

Harackiewicz, J.M., Manderlink, G. and Sansone, C. (1992) 'Competence processes and achievement motivation: implications for intrinsic motivation', in A.K. Boggiano and T.S. Pittman (Ed.s) *Achievement and Motivation: a Social-Developmental Perspective*. U.K.: Cambridge University Press.

Hedges, L.V. and Olkin, I. (1980) 'Vote counting methods in research synthesis'. *Psychological Bulletin*, 88, 2, 359-69.

Hedges, L.V. and Olkin, I. (1985) *Statistical Methods for Meta-Analysis*. Orlando, Florida: Academic Press.

Hill, P.W. and Rowe, K.J. (1996) 'Multilevel modelling in school effectiveness research'. *School Effectiveness and School Improvement*, 7, 1, 1-34.

Holland, P.W. (1986) 'Statistics and causal inference'. *Journal of the American Statistical Association*, 81, 945-71.

Hopkins, D. and Lagerweij, N. (1996) 'The school improvement knowledge base' in D. Reynolds, R. Bollen, B. Creemers, D. Hopkins, L. Stoll and N. Lagerweij (eds.) *Making Good Schools: Linking School Effectiveness and School Improvement*. London: Routledge.

Hull, C. (1985) 'Between the lines: The analysis of interview data as an exact art', *British Educational Research Journal*, 11, 1, 27-31.

Hunter, J.E. and Schmidt, F.L. (1996) 'Cumulative research knowledge and social policy formulation: The critical role of meta-analysis' *Psychology Public Policy and Law*, 2 (2) 324-347.

Ilgen, D.R., Fisher, C. D. and Taylor, M.S. (1979) 'Consequences of individual feedback on behavior in organizations'. *Journal of Applied Psychology*, 64, 349-71.

Jussim, L., Yen, H. and Aiello, J.R. (1995) 'Self-consistency, self-enhancement and accuracy in reactions to feedback'. *Journal of Experimental Psychology*, 31, 322-356.

Kaiser, H.F. (1974) 'An index of factorial simplicity'. *Psychometrika*, 39, 31-36.

Kerlinger, F.N. (1986) *Foundations of Behavioral Research*. New York: Holt, Rinehart and Winston.

Kluger, A.N. and DeNisi, A. (1996) 'The effects of Feedback Interventions on performance: a historical review, a meta-analysis, and a preliminary Feedback Intervention Theory'. *Psychological Bulletin*, 119, 2, 254-284.

Kulik, J.A. and Kulik, C.C. (1988) 'Timing of feedback and verbal learning.' *Review of Educational Research*, 58, 1, 79-97.

Leo, E.L., and Galloway, D. (1996). 'Evaluating research on motivation: generating more heat than light?'. *Evaluation and Research in Education*, 10, 1, 35-47.

Lepper, M.R., Greene, D., and Nisbett, R.E. (1973) 'Undermining children's intrinsic interest with external reward: a test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, 26, 129-37.

Levine, D.K. and Lezotte, L.W. (1990) *Unusually Effective Schools: a Review and Analysis of Research and Practice*. Madison, Wisconsin: National Center for Effective Schools Research and Development.

Locke, E.A., and Latham, G.P. (1984) *Goal Setting: a Motivational Technique that Works*. Englewood Cliffs, NJ: Prentice Hall.

Locke, E.A., and Latham, G.P. (1990) *A Theory of Goal Setting and Task Performance*. Englewood Cliffs, NJ: Prentice Hall.

Locke, E.A., Saari, L.M., Shaw, K.N. and Latham, G.P. (1981) 'Goal setting and task performance 1969-1980'. *Psychological Bulletin*, 90, 125-152.

Lyakowski, R.S. and Walberg, H.J. (1982) 'Instructional effects of cues, participation, and corrective feedback: a quantitative synthesis.' *American Educational Research Journal*, 19, 4, 559-78.

Madaus, G.F., Kellaghan, T., Rakow, E.A. and King, D.J. (1979) 'The sensitivity of measures of school effectiveness'. *Harvard Educational Review*, 49 (2) 207-230.

Maehr, M.L. (1983) 'On doing well in science: Why Johnny no longer excels; why Sarah never did' in S.G. Paris, G.M. Olson and H.W. Stevenson (Ed.s), *Learning and Motivation in the Classroom*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Marsh, H.W. (1990) 'Causal ordering of academic self-concept and academic achievement: a multiwave, longitudinal panel analysis'. *Journal of Educational Psychology*, 82 (4) 646-656.

McClelland, D.C. (1961) *The Achieving Society*. Princeton: Van Nostrand.

McColskey, W. and Leary, M.R. (1985) 'Differential effects of norm-referenced and self-referenced feedback on performance expectancies, attributions and motivation'. *Contemporary Educational Psychology*, 10, 275-284.

McKennell, A.C., (1977) 'Attitude Scale Construction' in C.A. O'Muircheartaigh and C. Payne (eds), *The analysis of survey data: Volume 1 Exploring Data Structures*. Chichester, UK: John Wiley and Sons.

Mesch, D.J., Farh, J-L. and Podsakoff, P.M. (1994) 'Effects of feedback sign on group goal setting, strategies and performance'. *Group and Organization Management*, 19, 3, 309-333.

Moreland, R.L. and Sweeney, P.D. (1994) 'Self-expectancies and reactions to evaluations of personal perfomance'. *Journal of Personality*, 52, 156-176.

Mortimore, P. (1991) 'The front page or yesterday's news: the reception of educational research' in G. Walford (Ed.) *Doing Educational Research*. London: Routledge.

Mortimore, P., Sammonds, P., Stoll, L., Lewis, D., and Ecob, R., (1988) *School Matters: The Junior Years*. Wells: Open Books.

Mory, E.H. (1992) 'The use of informational feedback in instruction: Implications for future research'. *Educational Technology Research and Development*, 40, 3, pp5-20.

Nicholls, J.G. (1983) 'Conceptions of ability and achievement motivation: a theory and its implications' in S.G. Paris, G.M. Olson and H.W. Stevenson (Ed.s), *Learning and Motivation in the Classroom*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Norusis, M.J. (1985) *SPSS^x Advanced Statistics Guide*. Chicago: SPSS Inc.

Oakes, M. (1986) *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.

Oppenheim, A.N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement* (New edition.). London: Pinter.

Plowden Report (1967) *Children and Their Primary Schools*. London: HMSO.

Podsakoff, P.M. and Farh, J-L. (1989) 'Effects of feedback sign and credibility on goal setting and task performance'. *Organizational Behavior and Human Decision Processes*, 44(1), 45-67.

Preece, P., (1989) 'Pitfalls in research on school and teacher effectiveness'. *Research Papers in Education*, 4 (3) 47-69.

Pritchard, R.D., Jones, S.D., Roth, P.L., Stuebing, K.K., and Ekeberg, S.E. (1988) 'Effects of group feedback, goal setting, and incentives on organizational productivity'. [Monograph], *Journal of Applied Psychology*, 73, 337-358.

Purkey S.C. and Smith, M.S. (1983) 'Effective schools: a review'. *The Elementary School Journal*, 4, 74-85.

Quadling, D. (1987) *Statistics and Probability*. Cambridge: Cambridge University Press.

Ralph, J.H. and Fennessey, J. (1983) 'Science or reform: some questions about the effective schools model', *Phi Delta Kappan*, 64, 10, 689-694.

Raudenbush, S.W. (1989) 'The analysis of longitudinal, multilevel data'. *International Journal of Educational Research*, 13, 721-740.

Reynolds, D. (1992) 'School effectiveness and school improvement: an updated review of the British literature', in D. Reynolds and P. Cuttance, *School Effectiveness: Research, Policy and Practice*. London: Cassell.

Reynolds, D. and Stoll, L. (1996) 'Merging school effectiveness and school improvement: the knowledge bases' in D. Reynolds, R. Bollen, B. Creemers, D. Hopkins, L. Stoll and N. Lagerweij (eds.) *Making Good Schools: Linking School Effectiveness and School Improvement*. London: Routledge.

Reynolds, D., Bollen, R., Creemers, B., Hopkins, D., Stoll, L. and Lagerweij, N. (1996) *Making Good Schools: Linking School Effectiveness and School Improvement*. London: Routledge.

Reynolds, D., Philips, D. and Davie, R. (1989) 'An effective school improvement programme based on school effectiveness research' *International Journal of Educational Research*, 13 (7) 801-14.

Rogosa, D., Brandt, D. and Zimowski, M. (1982) 'A growth curve approach to the measurement of change', *Psychological Bulletin*, 92, 3, 726-48.

Rosenthal, R. (1979) 'The "file drawer problem" and tolerance for null results' *Psychological Bulletin*, 86, 638-641.

Rotter, J.B. (1966) 'Generalized expectancies for internal versus external control of reinforcement.' *Psychological Monographs*, 80 (1, Whole No. 609).

Rutter, M., Maughan, B., Mortimore, P., and Ouston, J., (1979) *Fifteen Thousand Hours*. London: Open Books.

Sammons, P., Hillman, J. and Mortimore, P. (1994) *Key Characteristics of Effective Schools: A Review of School Effectiveness Research*. London: OFSTED.

Sammons, P., Mortimore, P., and Thomas, S. (1996) 'Do schools perform consistently across outcomes and areas?' in J. Gray, D. Reynolds, C. Fitz-Gibbon and D. Jesson, *Merging Traditions: The Future of Research on School Effectiveness and School Improvement*. London: Cassell.

Savitz, D.A. (1993) 'Is statistical significance testing useful in interpreting data?' *Reproductive Toxicology*, 7, 2, 95-100.

SCAA (School Curriculum and Assessment Authority) (1994) *Value added performance indicators for schools*. London: SCAA.

Scheerens, J. (1992) *Effective Schooling: Research, Theory and Practice*. London: Cassell.

Scheerens, J. and Creemers, B. (1989) 'Towards a more comprehensive conceptualization of school effectiveness', in B. Creemers, T. Peters and D. Reynolds (eds.) *School Effectiveness and School Improvement: Proceedings of the Second International Congress, Rotterdam, 1989*. Amsterdam: Swets and Zeitlinger.

Shaver, J.P. (1993) 'What statistical significance testing is, and what it is not'. *Journal of Experimental Education*, 61, 4, 293-316.

Simon, H.A. (1974) 'How big is a chunk?' *Science*, 183, 482-8.

Slavin, R.E. (1980) 'Effects of individual learning expectations on student achievement'. *Journal of Educational Psychology*, 72, 4, 520-4.

Teddlie, C. and Stringfield, S. (1993) *Schools make a Difference: Lessons Learned from a 10-year Study of School Effects*. Teachers College Press: London.

Thompson, B. (1992) 'Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development*, 70, 434-438.

Thompson, B. (1994) 'The pivotal role of replication in psychological research: empirically evaluating the replicability of sample results'. *Journal of Personality*, 62, 2, 157-176.

Thompson, B. (1996) 'AERA editorial policies regarding statistical significance testing: three suggested reforms'. *Educational Researcher*, 25, 2, 26-30.

Trower, P. and Vincent, L. (1995) *The Value Added National Project: Technical Report, Secondary*. London: SCAA.

Tukey, J.W. (1969) 'Analyzing data: Sanctification or detective work?' *American Psychologist*, 24, 83-91.

Tukey, J.W. (1977) *Exploratory data analysis*. Reading MA: Addison-Wesley.

Tymms P.B. (1990) 'Can indicator systems improve the effectiveness of science and mathematics education? The case of the UK', *Evaluation and Research in Education*, 4, 2, 61-73.

Tymms P.B. (1995) 'Influencing educational practice through performance indicators.' *School Effectiveness and School Improvement*, 6, 2, 123-145.

Tymms, P. (1996) 'Theories, Models and Simulations: School effectiveness at an impasse', in J. Gray, D. Reynolds, C. Fitz-Gibbon and D. Jesson, *Merging Traditions: The Future of Research on School Effectiveness and School Improvement*. London: Cassell.

Tymms, P. (1997a) 'Responses of headteachers to value added and the impact of feedback' The Value Added National Project, Technical Report: Primary 3. London: SCAA.

Tymms, P. (1997b) 'The impact of indicators on primary schools.' Paper presented at Evidence-Based Policies and Indicator Systems Conference, University of Durham, July 1997.

Tymms, P. and Fitz-Gibbon, C.T. (1990) 'The stability of school effectiveness indicators' Moray House, Edinburgh and CEM, School of Education, University of Newcastle upon Tyne..

Van Ryckeghem, M. and Brutten, G.J. (1992) 'The Communication Attitude Test: A test-retest reliability investigation'. *Journal of Fluency Disorders*, 17, 3, 177-190.

Viswanatham, M. (1994) 'On the test-retest reliability of the Preference for Numerical Information scale'. *Psychological Reports*, 75, 1(1), 285-286.

Waldersee, R., and Luthans, F. (1994) 'The impact of positive and corrective feedback on customer service performance'. *Journal of Organizational Behavior*, 15, 83-95.

Walford, G. (1991) 'Reflexive accounts of doing educational research' in G. Walford (ed.), *Doing Educational Research*. London: Routledge.

Weiner, B. (1972) 'Attribution theory, achievement motivation, and the educational process.' *Review of Educational Research*, 42, 2, 203-215.

Weiner, B. (1992) *Human Motivation: Metaphors, Theories and Research*. Newbury Park, CA: Sage.

Willms, J.D. (1992) *Monitoring School Performance: a Guide for Educators*. London: Falmer.

Wood, R.E., Mento, A.J. and Locke, E.A. (1987) 'Task complexity as a moderator of goal effects: a meta-analysis'. *Journal of Applied Psychology*, 72, 416-25.

Zimmerman, B.J., Bandura, A. and Martinez-Pons, M. (1992) Self-motivation for academic attainment: the role of self-efficacy beliefs and personal goal setting'. *American Educational Research Journal*, 29 (3) 663-676.