Frederick John McCausland

TEACHERS' PREDICTIONS AND PUPIL ACHIEVEMENT
IN SCHOOL CERTIFICATE

Submitted for the
degree of Master of Arts in Education
at the
Victoria University of Wellington
1981

ABSTRACT

This thesis examines the estimates of pupils' performances in School Certificate made by teachers at Hutt Valley High School over a six year period. It aims to assess the degree of accuracy of teacher predictions, in particular at the pass/fail boundary, and to identify and evaluate the relative importance of some correlates of accurate prediction.

The data for this study is taken from school records showing for each class the name of the teacher and the estimated and achieved marks of each pupil in School Certificate for the period 1973 to 1978.

The statistical analyses include the calculation of the Pearson Product-Moment Correlation Coefficient between the estimated and achieved marks of pupils in English, Mathematics, Science and French. Expectancy tables are provided to give greater detail of the accuracy of teacher's estimates at different points across the mark range. A multiple regression analysis is carried out in order to identify and evaluate the relative importance of some correlates of accurate prediction. The statistical section concludes with a summary analysis of the data from the subjects for the year following the major analysis. The results of the study show that the correlations between the estimated and achieved marks are greater in Mathematics and French than in English and Science, however, the estimated and achieved class means across subjects show greatest agreement in English and least agreement in Mathematics. Whilst variations across subjects and across years occur, in general teachers show a tendency to underestimate pupils at the upper end of the ability range and overestimate the marks of pupils at other points of the scale. A movement of approximately 16.5% across the pass/fail boundary was observed when comparing the estimated and achieved marks of pupils. The study also concludes that

the only factor to reveal a substantial and systematic link with the quality of teachers' predictions was the ability level of the class under analysis - teachers were generally more able to predict the results of their pupils for higher ability than lower ability classes (English excepted).

ACKNOWLEDGMENTS

The author is grateful to Dr C.G.W. Hall for his interest in the development of this thesis and in particular for his advice on the literature which related to the topic. Thanks are also due to Professor R.W. Marsh for his interest and advice. The encouragement of Paul Canham, Barbara Mabbett and other friends for their criticism of the text is very much appreciated. Acknowledgement is given to Victoria University for the use of computer facilities.

CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	iii
CONTENTS	iv
LIST OF TABLES	v
LIST OF FIGURES	vii
CHAPTER ONE	
Introduction	1
1.1 Background and general aims	1
1.2 Plan of the study	3
CHAPTER TWO	
Review of the literature	4
2.1 Introduction	4
2.2 Research overseas	5
2.3 Research in New Zealand	27
2.4 Summary	34
CHAPTER THREE	
Data and statistical analyses	36
3.1 Introduction	36
3.2 Data	36
3.3 Statistical analyses	38
CHAPTER FOUR	
Presentation of results	41
4.1 Introduction	41
4.2 Correlation between the estimated and achieved marks of pupils in School Certificate for each class	41
4.3 Expectancy table analysis of the estimated and achieved marks of pupils in School Certificate for each year	56
4.4 Multiple regression analysis of the accuracy of teachers' estimates	90
4.5 Summary analysis and comparison of data for 1977 and 1978	102
CHAPTER FIVE	
Discussion, interpretation and integration	105
5.1 Introduction	105
5.2 Discussion and interpretation of results	121
5.3 Summary of findings	124
BIBLIOGRAPHY	

LIST OF TABLES

2.1	Summary of correlations obtained by Petch relating teachers' estimates to pupils' attainment	9
2.2	Summary of findings by Sowell of the predicted and achieved grades of pupils in one CSE Board	16
2.3	Range of correlations obtained by Murphy (1979) for the estimated and achieved rank orders of pupils at GCE	17
2.4	Comparison of the predicted and achieved grades for five different groups of examination candidates	19
2.5	Agreement between school assessments and School Certificate in eight subjects - from Hulbert (1978)	29
2.6	Summary of correlations obtained by Gilmore between school assessments and School Certificate marks	31
4.1a	Mean scores and correlations for English 1973	43
4.1b	Mean scores and correlations for English 1974	43
4.1c	Mean scores and correlations for English 1975	44
4.1d	Mean scores and correlations for English 1976	44
4.1e	Mean scores and correlations for English 1977	45
4.2a	Mean scores and correlations for Mathematics 1973	46
4.2b	Mean scores and correlations for Mathematics 1974	46
4.2c	Mean scores and correlations for Mathematics 1975	47
4.2d	Mean scores and correlations for Mathematics 1976	47
4.2e	Mean scores and correlations for Mathematics 1977	48
4.3a	Mean scores and correlations for Science 1973	49
4.3b	Mean scores and correlations for Science 1974	49
4.3c	Mean scores and correlations for Science 1975	50
4.3d	Mean scores and correlations for Science 1976	50
4.3e	Mean scores and correlations for Science 1977	51
4.4a	Mean scores and correlations for French 1973	52
4.4b	Mean scores and correlations for French 1974	52
4.4c	Mean scores and correlations for French 1975	52
4.4d	Mean scores and correlations for French 1976	52
4.4e	Mean scores and correlations for French 1977	52
4.5	Mean and median correlations between the estimated and achieved scores of pupils by subject for each year	53
4.6a	Grouped frequency (expectancy) table relating estimated mark to achieved mark, English 1974	58
4.6b	Grouped frequency (expectancy) table relating estimated mark to achieved mark, English 1975	59
4.6c	Grouped frequency (expectancy) table relating estimated mark to achieved mark, English 1976	60
4.6d	Grouped frequency (expectancy) table relating estimated mark to achieved mark, English 1977	61

4.7a	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Mathematics 1974	62
4.7b	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Mathematics 1975	63
4.7c	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Mathematics 1976	64
4.7d	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Mathematics 1977	65
4.8a	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Science 1974	66
4.8b	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Science 1975	67
4.8c	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Science 1976	68
4.8d	Grouped frequency (expectancy) table relating estimated mark to achieved mark, Science 1977	69
4.9a	Grouped frequency (expectancy) table relating estimated mark to achieved mark, French 1974	70
4.9b	Grouped frequency (expectancy) table relating estimated mark to achieved mark, French 1975	71
4.9c	Grouped frequency (expectancy) table relating estimated mark to achieved mark, French 1976	72
4.9d	Grouped frequency (expectancy) table relating estimated mark to achieved mark, French 1977	73
4.10a	Summary of movement over pass/fail boundary	74
4.10b	Summary of percentages of pupils over and underestimated 1974-77	77
4.11	Relating the estimated number of subject passes to the achieved number of subject passes in 1974	85
4.12	Relating the estimated number of subject passes to the achieved number of subject passes in 1975	86
4.13	Relating the estimated number of subject passes to the achieved number of subject passes in 1976	87
4.14	Relating the estimated number of subject passes to the achieved number of subject passes in 1977	88
4.15	Summary of movement over 6th form entry boundary	89
4.16	Correlation matrix for English 1973-1977 combined	92
4.17	Correlation matrix for Mathematics 1973-1977 combined	94
4.18	Correlation matrix for Science 1973-1977 combined	96
4.19	Correlation matrix for French 1973-1977 combined	98
4.20	Multiple regression analysis of the accuracy of teachers' estimates of pupils' marks by subject, 1973 to 1977 combined.	100
4.21	Summary of the analysis of the estimated and achieved marks of pupils in 14 subjects for 1977 and 1978	103

LIST OF FIGURES

4.1a	Frequency histogram for English for the years 1974 to 1977 combined.	78
4.1b	Cumulative frequency distribution for English for the years 1974 to 1977 combined.	78
4.2a	Frequency histogram for Mathematics for the years 1974 to 1977 combined.	79
4.2b	Cumulative frequency distribution for Mathematics for the years 1974 to 1977 combined.	79
4.3a	Frequency histogram for Science for the years 1974 to 1977 combined.	80
4.3b	Cumulative frequency distribution for Science for the years 1974 to 1977 combined.	80
4.4a	Frequency histogram for French for the years 1974 to 1977 combined.	81
4.4b	Cumulative frequency distribution for French for the years 1974 to 1977 combined.	81
4.5a	Frequency histogram for English, Mathematics, Science, French, for the years 1974 to 1977 combined.	82
4.5b	Cumulative frequency distribution for English, Mathematics, Science, French for the years 1974 to 1977 combined.	82

*Chapter One*INTRODUCTION1.1 Background and general aims

New Zealand State Secondary Schools need to know by about November each year, the likely enrolment figures for the next year. The appropriate staffing of each subject area requires a fairly accurate projection of the composition of all classes at every level from the third form through to the seventh form. Fourth and seventh form rolls can be compiled with a reasonable degree of certainty from information gathered directly from the pupils. Third form details are obtained from information supplied by the contributing primary schools. Fifth and sixth form classes are much more difficult to project because the School Certificate examination results are not released until towards the end of January - results determine the final composition of senior classes.

While it is usually possible to calculate the number of pupils likely to qualify for entry to the sixth form, it is not so easy to determine the numbers of pupils who wish to take each subject. In order to reduce the degree of uncertainty of the projected fifth and sixth form rolls, Hutt Valley High School has, over a period of years, developed a system of estimating pupils' School Certificate results. The advantage of such a system is that it enables the people responsible for the forward planning to make an estimate of the number of pupils likely to be studying in each subject area. This in turn makes it much easier to have the correct number of staff available to take the classes. Accurate estimates also help to reduce the chance of forming excessively large or unnecessarily small classes.

After the end-of-year examinations have been completed and fifth form reports compiled, teachers of fifth form classes are asked to

estimate the mark which they expect each of their pupils to gain in the School Certificate examination. When making these estimates it has been found in practice that teachers rely substantially on their professional judgement. They take into account not only performance in school examinations and tests, but also a pupil's attitude, work throughout the year, temperament and motivation. Less experienced teachers usually confer with their colleagues and compare their tentative estimates with those of a parallel class. The information is then entered into a record book into which is subsequently placed the achieved mark in School Certificate.

With the help of the record book it has been possible to make a reasonable estimate of the number of pupils likely to be in the fifth and sixth form at the beginning of the following year. The school roll is now over 1400 so that a very small percentage variation in the number of passes in School Certificate can alter the staffing ratio; for example, twenty more pupils in the senior school allows for the appointment of a further full-time teacher while a variation of 20 pupils between the estimated and actual rolls could mean that some classes are either too big or too small, or even that a class is without a suitably qualified teacher. The present study was devised in the hope that better understanding of teachers' estimation of School Certificate results would result in more reliable forecasts, which would in turn produce more effective planning.

The carefully recorded estimates made by teachers, the actual marks obtained by pupils, and certain hypothesized relevant background details of teachers (for example, length of experience and qualifications) and their classes comprise the data for this study. In general terms, the major aims of this thesis are:

- (1) to assess the degree of accuracy of teachers' predictions, in particular at the pass/fail boundary, and

- (2) to identify and evaluate the relative importance of some correlates of accurate prediction.

1.2 Plan of the study

The present chapter has focussed, albeit briefly, on the background and aims of this research; the next chapter contains a detailed coverage of the literature which highlights, in particular, evidence of relevance to the first of the preceding aims. Chapter Three provides further details of the data collected by (and/or available to) the writer; a full description of the statistical analyses employed is also given. Chapter Four provides a full presentation and description of the results - interpretations and discussion are kept to a minimum. Lastly, Chapter Five expands upon the results - interpretations are suggested and, where relevant, the present results are integrated with findings from the review of the literature in Chapter Two; Chapter Five also contains suggestions for further research and provides at the end a short summary of the major findings of this research.

*Chapter Two*REVIEW OF THE LITERATURE2.1 Introduction

The review of the literature is presented under two major headings dealing with research carried out overseas (Section 2.2) and research carried out in New Zealand (Section 2.3). The context for some of the research to be reviewed is the movement both overseas and in New Zealand to a school based system of assessing performance in public examinations (e.g. School Certificate in New Zealand and Mode III of the Certificate of Secondary Education in Britain). In addition, evidence is provided from research dealing with procedures for selecting pupils for secondary education in England and Scotland (e.g. the 11-plus examination). It should be stated that it is not the objective of this thesis nor this review to discuss the merits or otherwise either of internal (school) assessment of secondary examinations or of the past (and in the case of one or two Local Education Authorities, the present) British policy of selecting pupils for different forms of secondary education. The point being made is that much of the research to be reviewed was carried out in a wider framework or context and with different aims and objectives than that of the present research; consequently, in a few instances importance is attached here to a finding which, in the context of the original research carried out, was dealt with in no more than one or two pages of a major report or text. Furthermore, in one important area, that of the identification of variables related to the ability of teachers to predict their pupils' results in external examinations, the writer has been unable to find a single research study which either controls for, or systematically investigates, what might be the important variables involved (e.g. age of the teacher, experience, qualifications, the ability of pupils

and so on). Finally, the comment should be made that the seemingly narrow scope of the review, focussing primarily on research carried out in Britain and New Zealand, reflects the comparable nature of many of the issues in both systems as related to the role of the teacher in assessing his or her pupils for the awards of certificates in public examinations.

2.2 Research Overseas

Evidence from the entrance examination to secondary school in Britain

"203 The results of research and the experience of examining bodies show that teachers are, as a rule, very accurate in placing their pupils in an order of merit; indeed we are satisfied that in this respect the teacher's grading is more trustworthy than any other.

204. But, if there is to be a uniform pass standard throughout the country, the marks submitted by different schools must be compared and equated, and it is here that difficulty arises, since teachers' estimates vary (1) in standard, and (2) in the spread of marks - a two-fold defect which, though in no way reflecting on the teachers' competence, does invalidate the raw marks as a basis for the award of a national certificate."

Scottish Education Department (1947),
paragraphs 203 and 204.

The major conclusion overall from the research to be reported, as well summarised in the above abstract from the Scottish Education Department, is that teachers are generally able to rank their pupils in order of merit but unable to define the level or standard of attainment of their class relative to other groups or to a national standard. Unfortunately, details of the "research and experience" referred to by the Scottish Education Department are not presented, nor is the comment

that "Teacher's grading is more trustworthy than any other" supported with figures or by specific reference to other research. However, a later study in England by Yates and Pidgeon (1957) cites evidence pointing to the usefulness of primary teachers' estimates in predicting subsequent success at secondary school. In particular, they draw upon evidence from McClelland (1942), Sutcliffe and Canham (1945)*, Bosomworth (1953)* and Richardson (1956) concerning the effectiveness of various criteria in predicting future attainment. The study by Richardson indicated that primary school assessments in the form of a rank order, pooled and scaled from four schools, provided the best single prediction with performance after one year at secondary school ($r = 0.835$), corrected for selectivity), that is, better than either a verbal intelligence measure ($r = 0.826$) or actual marks in the local authority entrance examinations in English ($r = 0.741$) and arithmetic ($r = 0.792$). Correlations for each of the rank orders by schools were not provided. The studies by Sutcliffe and Canham (1945) and Bosomworth (1953) similarly indicated that, suitably scaled, teachers' assessments yield as good a prediction of future attainment as the traditional public examinations. Yates and Pidgeon also report that the Bosomworth study, in particular, showed that teachers estimates were only as good as other measures in predicting attainment if they had been suitably scaled; without scaling the assessments were of little value. In contrast to these studies, McClelland (1942, 1949) found that teachers' estimates were less effective than either a battery of intelligence tests (Moray House) or external examinations of the traditional style in English and arithmetic. However, the teachers' estimates were still included with the other predictors in yielding in combination

* The writer has been unable to access these sources at first hand.

the most satisfactory prediction overall. More directly of relevance to the present research, McClelland also reported correlations between the estimated and achieved percentages of pupils from twelve schools for the entrance examinations in English and arithmetic; in addition he compared the mean scores and standard deviations for each set of marks. The correlations for English ranged from 0.94 to 0.60 with a median value of 0.83; in arithmetic the correlations ranged from 0.88 to 0.61 with a median value of 0.77. Taken as a whole, these values indicate at least a moderate agreement between the estimated and achieved marks of pupils with respect to the order of pupils; however, a comparison of the mean scores shows that in each and every school students were expected to perform better than their resultant entrance marks indicated. In English, a weighted mean difference overall of 15.0 per cent was obtained; in arithmetic the corresponding figure was 24.8 per cent. In summarising these results, McClelland concluded that the estimates were generally too high although with a tendency for teachers to underestimate the performances of the good pupils; he also concluded that the standard (leniency-severity) of the estimates varied considerably from school to school.

In summarising their review of research, Yates and Pidgeon conclude:

"iv) Teachers' estimates of their pupils' capabilities can add to the predictive value of a battery of tests or examinations, but only if these estimates are appropriately scaled to allow for differences between the levels of performance of children from different schools. Whether or not teachers' estimates provide a better prediction than tests or examinations is not clear from the researches that we have examined although there are indications to suggest that, if the problem of scaling is satisfactorily solved, teachers' estimates may well be found

to be as effective as, if not more effective than the objective tests or examinations."

(p.47)

In their own study, Yates and Pidgeon report correlations between the assessments of pupils made by the Heads of 19 primary schools and the pupils' total scores (English and arithmetic) in the entrance examination. The values obtained ranged from 0.99 to 0.76 with a median value of 0.92 - again indicating a good agreement in terms of the order of merit of pupils. Yates and Pidgeon also report correlations between the assessment of primary heads (pooled and scaled across schools) and attainment after two and three years of secondary schooling; in both cases the values exceeded those obtained from several standardised ability measures (e.g. verbal, English, arithmetic and spatial). A follow-up study of borderline entrance cases, using an achievement criteria based on secondary teachers' assessments scaled with reference to a battery of ability measures, also revealed that the primary heads' assessments contained fewer 'misfits', that is, errors in placing borderline students than all but the entrance examination itself (approximately 28 per cent in the case of the assessments and 27 per cent in the case of the examination). It should be added that the method of selection for secondary schooling adopted by the education authority, a composite of several measures, provided around nine per cent fewer errors than either the entrance examination or the primary heads' assessments.

Research related to assessment for CSE and GCE

To date there have been three major published studies in Britain each with the primary aims of determining the level of agreement between the assessment of pupils made by teachers and the subsequent performances of pupils in external (public) examinations. The first of

these, a study at GCE Advanced Level (the 18-plus age group), represents perhaps the most extensive work in the area. Petch (1964) compared the estimated and achieved marks of 101 different groups of pupils in 12 subjects; schools contributing assessments were selected at random from the intake for one examining board subject to the provision that the school entered at least 20 pupils in the popular mainstream subjects under study and 10 pupils in the less popular subjects. A summary of the correlations obtained by Petch is presented in Table 2.1 (extracted from Tables 6-9 of Appendix A in Petch) and shows the number of schools contributing estimates, the highest and lowest correlation by school for each subject, and the median correlation by school for each subject. Subjects are ordered in terms of the size of the median correlation.

Table 2.1 Summary of correlations obtained by Petch relating teachers' estimates to pupils' attainment

Subject	No. of Schools	Highest	Correlation	
			Lowest	Median
German	9	0.97	0.75	0.91
Mathematics	11	0.95	0.70	0.91
Further Maths	9	0.92	0.71	0.87
Chemistry	7	0.89	0.63	0.84
French	7	0.96	0.78	0.82
Latin	9	0.84	0.59	0.79
Physics	8	0.84	0.69	0.76
Biology	10	0.85	0.35	0.73
English Lit.	8	0.75	0.50	0.72
History	8	0.92	0.56	0.70
Geography	10	0.81	0.45	0.69
Zoology	5	0.77	0.58	0.66

It is evident from Table 2.1 that the level of agreement between the estimates and attainments of pupils differs widely both between and within subjects. It is also apparent that the assessments in Mathematics and the Languages resembled more closely, at least in order of merit, the examination marks of pupils than the assessments in art subjects

(English, History and Geography) and science subjects with a greater 'verbal' content (Biology and Zoology). These results are perhaps not surprising in that the subjects in the upper half of the table generally provide a more objective criteria for assessing attainment than the subjects in the lower half. It could be argued, in fact, that these results pre-empt the more recent evidence from reliability studies of British examinations (Willmott and Hall (1975) and Willmott and Nuttall (1975)) which show higher estimates in general for Mathematics, the Languages and certain science subjects (Physics and Chemistry) than for English, History, Religious Studies and the various craft subjects (e.g. Woodwork and Metalwork). From the viewpoint of validity, however, it cannot be automatically concluded that the subjects in the upper half of Table 2.1. have yielded a more satisfactory set of results; as Petch himself points out:

"Studying at school is not, or may not be, wholly directed towards passing examinations. The qualities tested in the examination room may not be the identical qualities which may emerge during the course of preparation." (p.6)

A similar viewpoint is expressed by the Schools Council (1966a) in a study designed to compare different methods of marking English composition which incorporated the use of teachers' assessments as a criterion of validity; the Council concluded:

"In this investigation the teachers' assessments were of little value. This was partly due to the fact that the instructions were comparatively open ... also in part due to the fact that a teacher's estimate is an assessment of a different kind from that made by examiners marking answers to an examination paper. The former is a more general and comprehensive assessment and the latter highly specific." (para. 133)

The finding that the estimates of teachers were of little value is discussed later; the comment that the assessment represents a more comprehensive evaluation of pupil learning than an external examination clearly points to the difficulty in trying to draw conclusions directly from Table 2.1. It should be added, however, that a moderate correlation at least ought to be forthcoming in that any content valid examination should represent the underlying coursework upon which it is based.

As Petch himself points out:

"But if the two orders (the teacher's assessments and the examiner's markings*) show no agreement whatsoever, it would seem reasonable to conclude either that the preparation has been unwise, even ill-advised, or that the examining has been incompetent or unfair, or at least that preparation and examining are out of joint." (p.6).

Perhaps all that can be said in the absence of a detailed description of the bases upon which teachers arrive at their assessments, is that a particular value is high, low or moderate bearing in mind the range of values typically obtained both in the study concerned and in other studies of a similar nature. If this rather loose criterion is applied to Table 2.1., it can be argued that the estimates provided by a few schools, particularly for those subjects in the lower half of the table, are somewhat 'out of joint' with the markings of examiners. This, at least, is the interpretation also drawn by Petch although not from applying (explicitly anyway) the criterion suggested above.

The study by Petch provides two further pieces of relevant information for the present research. Firstly, the question of standards or scaling was investigated by comparing the grades of pupils (a 5-point scale) as awarded by the examining board with the grades

* Inserted by the writer.

corresponding to the predicted scores given by teachers. Whilst a breakdown by each of the subjects was not provided, an analysis of the results overall (all subjects pooled) showed a clear tendency for teachers to overestimate more often their pupils results than to underestimate them. For example, in terms of passes and failures, teachers correctly predicted the results for 78.6 per cent of pupils; of the remaining 21.4 per cent of pupils 18.5 per cent were expected to pass but in fact failed. Using the full five-point grade scale, 42.8 per cent of pupils obtained lower grades than expected whilst only 18.2 per cent obtained higher grades than expected. As will be seen later, these results agree fairly closely with similar analyses carried out on both GCE O-level and CSE examinations.

The other information of relevance to the present study concerned a follow-up by Petch of explanations from schools for the discrepancies between the assessments by teachers and the marks awarded by the examiners. In the main, explanations were sought for largely discrepant rankings (for the school group overall) and for extreme discrepancies in the case of individual pupils. It is not possible to report in detail here the many and varied comments given although the following brief extract from one letter indicates the two-sided nature of the discrepancies:

"We, on our side, were rather tempted to ask for your explanation of the results". (p.19)

Perhaps the major conclusion that can be drawn from the various comments given is that they affirm the broad base upon which assessments are made of pupils by teachers - at least in the case of 'discrepant' schools. Obviously this conclusion prompts the suggestion that any further research should include a comparison of the bases given for the assessments by teachers in both 'discrepant' and

'non-discrepant' schools or even within a school between 'discrepant' and 'non-discrepant' pupils. (Unfortunately, much of the data in the present study was collected for school records well before this research was contemplated; it would be unreasonable to expect teachers to recall the basis of their assessments in classes going back three or four years in time, indeed, several of the teachers are no longer at the school.) The following conclusion, in the form of an 'informal recommendation', perhaps summarises best the relative weights which Petch feels should be given to the external examinations and the assessments by teachers:

"A certificate setting out in one column the school's opinion which specifically excluded any element of forecast of the probable examination performance and, in a second column, the examiners' verdict, arrived at without any knowledge of the school estimate, would give a better, because a more rounded, view of the pupil/candidate than either assessment gives independently." (p.29)

One final point should be made; in his study Petch carefully instructed teachers not to make actual forecasts or predictions of how pupils will fare in A-level but rather to assess them in terms of effort and performance in class. In his own words ...

"a difference can be seen between a school's estimate or appraisal stating what, in its opinion, pupils 'are worth' (SC. in the form-room) and the school's estimate or conjecture of what they 'will get' (SC. in the examination room), the latter more regardless or less regardless of what they 'ought to get' on grounds other perhaps than actual examination performance. The difference however is not absolute: it is rather a matter of emphasis." (p.5)

Whilst *prima facie* this distinction may have appeal, no actual evidence is brought to bear which supports it even in 'emphasis' only. Indeed, in the experience of the writer, teachers take account of a wide range of factors when forecasting their pupils' examination results; for example, they appear to consider each pupil's performance in school examinations and tests, his or her attitude and work throughout the year, temperament and motivation, and even the views of colleagues in cases of doubt. In other words, the writer is uncertain as to whether the distinction made by Petch has a basis in practice even in 'emphasis' only. Further, it is interesting that neither of the two other major studies published in Britain to date make this distinction; Sowell (1970) specifically asked teachers to 'forecast' their pupils results at CSE whilst Murphy (1979) asked teachers to assess their pupils in the form of predicted O- and A-level GCE grades. It is interesting, as will be seen shortly, that they obtained very similar results to that of Petch.

The primary objective of the study by Sowell was to determine whether teachers were able to interpret the guidelines associated with the new CSE examinations in respect of the different grades awarded - in other words whether teachers were able to recognise from the criteria given the work of pupils worthy of Grade 1, Grade 2 and so on. (CSE comprises Grades 1 (high) to 5 (low) with a sixth category designated 'U' for work of a very low standard.) In addition, Sowell was interested to see if the interpretations of teachers improved with increased familiarity of the standards associated with CSE. To these ends Sowell compared the predicted and achieved grades of pupils in eight CSE subjects for the years 1965 and 1967; full details of the sampling procedure used were not provided although it appears that pupils were randomly chosen to provide sample sizes generally, but not in every case, of between 500 and 1,000 in each subject. His findings

for each subject are summarized in Table 2.2. below; it should be noted that whilst Sowell reports correlations overall (all subjects pooled) between the predicted and achieved grades of pupils, no values are presented separately for the subjects in the analysis. The overall correlations (0.698 for 1965 and 0.690 for 1967) agree closely with those computed by Sowell from the data used by Petch (0.679) and another CSE board (0.630)*; it should be added that the method used to compute the correlations (eta) is appropriate for coarsely graded distributions.

The figures in Table 2.2 point to two fairly clear conclusions: first, the mean grades reveal a good deal of optimism on the part of teachers in making their forecasts - in each year the predicted mean grade exceeded the achieved mean grade in all but one subject. Secondly, the figures for the percentage of accurate forecasts indicate that teachers of Homecraft and Metalwork yielded fewer errors than, in particular, teachers of Biology and Technical Drawing. Whilst the results in Table 2.2. show no evidence overall of an increase with time in the ability of teachers to make accurate forecasts, Sowell reports that fewer errors of prediction were made at the Grade 1 level in 1967, as compared to 1965, in all but French. The remaining conclusion reached by Sowell was that his results showed a marked similarity with the accuracy level of forecasts reported in two other studies (Petch and an unreferenced study); a comparison of these results is given shortly.

The third major study of the accuracy of teachers' assessments was that carried out by Murphy (1979) and involved three separate analyses of the O- and A-level results of pupils in one GCE examining

* not referenced by Sowell

Table 2.2 Summary of findings by Sowell of the predicted and achieved grades of pupils in
one CSE Board

Subject	1965				1967			
	Mean Grade:		% of forecasts		Mean Grade:		% of forecasts	
	Predicted	Achieved	Correct	Within one grade	Predicted	Achieved	Correct	Within one grade
Biology	2.86	3.27	32.2	79.8	3.04	3.62	33.2	77.7
English	2.96	3.13	41.1	88.5	2.94	3.07	44.1	91.3
French	3.32	3.27	43.4	89.0	2.96	3.20	41.0	85.0
History	2.99	3.27	42.6	85.7	3.09	3.18	41.8	86.9
Homecraft	2.68	2.95	42.2	93.4	2.67	2.68	52.6	97.2
Mathematics	3.08	3.27	41.5	85.0	3.11	3.13	41.7	82.9
Metalwork	2.76	2.92	46.0	93.8	2.81	2.78	45.3	90.7
Technical Drawing	2.70	3.18	33.4	76.4	2.87	3.12	35.2	79.2

board. The first analysis compared the predicted and achieved grades of 96 separate groups of candidates covering a wide range of subject examinations (Murphy, however, does not report results separately for each of the subjects in the analysis); in all, a total sample of 1,892 candidates was used comprising 1601 O-level and 291 A-level entries. His main finding from this analysis was a slight tendency overall for teachers to overestimate rather than underestimate their pupils' grades. Figures of 39.5, 80.4, 33.1 and 27.4 were obtained respectively for the percentages of correctly allocated grades, correctly allocated within one grade, the number of overestimates of grades and the number of underestimates of grades. An additional finding was that the results for A-level generally contained more errors of prediction than those for O-level.

For the second analysis carried out by Murphy, teachers were asked in advance of the examinations to rank their pupils in terms of their achievement in class throughout the year; these rankings were then correlated with the order of pupils as identified by the examination itself. A very wide variation in coefficients was in fact obtained for the 96 groups in the study - these are summarised below in Table 2.3.

Table 2.3. Range of correlations obtained by Murphy (1979) for the estimated and achieved rank orders of pupils at GCE

<u>Correlation</u>	<u>Number of Values</u>
0.90 +	3
0.80 - 0.89	20
0.70 - 0.79	16
0.60 - 0.69	22
0.50 - 0.59	13
0.40 - 0.49	13
0.30 - 0.39	4
< 0.30	5

Murphy reports an average value of 0.66 from the table given; clearly the general order of values obtained is lower than that found by Petch at A-level (the lowest median correlation by subject was in fact 0.66 for Zoology - see Table 2.1.). Murphy argues from these results that it would be unwise for examining boards to accept the rankings provided by teachers on the assumption that they provide similar results to those obtained from external examinations. In other words, he questions the general conclusion for a number of studies that teachers 'can rank but not scale' their pupils. The essential point made by Murphy is that the bases for the two systems of assessment - that from the school and that from the external examining body - may be quite different and need not automatically yield similar results. One could argue that these comments provide further support for the recommendation by Petch of a dual school and external examination certificate for pupils.

For the remaining analysis, Murphy asked teachers to comment directly on the scripts of pupils whose predicted grade was more than one point different from the grade awarded by the examiners. This involved sending the scripts of the 'discrepant' pupils back to the teachers who had made the original predictions - in all, it was possible to report the comments from 84 per cent of the cases identified (some teachers had moved). In general, higher than expected grades were explained by an unexpected improvement in performance, whereas lower than expected grades were explained either by an unexpected poor performance or a marked difference in standards between the teacher and examiners; in a few cases no explanation could be offered for the discrepancy. In summarising this section of his research, Murphy again emphasised the essential difference between the assessments made by teachers and those provided by the external examination, namely, the wider view of a candidate provided by the teacher as opposed to the

narrower but perhaps more objective view provided by the external examination.

The results of the study by Murphy are brought together with those by Petch, Sowell and the 'other' CSE board (referred to by Sowell) in Table 2.4. Figures are provided for the percentages corresponding in each study to the number of pupils whose grades were correctly predicted, the number whose grades were correct within plus or minus one grade, the number of pupils whose grades were overestimated and the number of pupils whose grades were underestimated (Table 2.4. is abbreviated from Table 3 in Murphy).

Table 2.4. Comparison of the predicted and achieved grades for five different groups of examination candidates

Study by:	Correct	Percentage of predictions		
		Within one grade	Over-estimated	Under-estimated
Petch	42.8	86.3	39.0	18.2
Sowell (1965)	40.0	85.3	37.1	22.9
Sowell (1967)	42.0	85.3	34.5	23.5
Other CSE Board	36.4	82.6	36.5	27.1
Murphy	39.5	80.4	33.1	27.4

It is evident from Table 2.4. that the various studies have yielded a fairly consistent picture of around 40 per cent for the number of correctly allocated grades, 84 per cent for the number of predictions correct within one grade, 36 per cent for the number of overestimates and 24 per cent for the number of underestimates. If the criterion of being correct within one grade is used, it can be reasonably argued that teachers by and large yeild similar results to the examiners with a tendency, on balance, to overestimate rather than underestimate their pupils' results.

Other British studies at the 16-plus and 18-plus age level

As indicated in the introduction to this review, a number of

studies have been carried out in Britain which, almost in passing, provide some evidence pointing to the usefulness (validity) of the estimates of pupil performance made by teachers. One such study, that by the Schools Council (1966a) has already been introduced; in this particular study the Schools Council experimented with three different marking procedures in order to obtain the best method for assessing English composition at O-level. They used as the main criterion of validity an independent assessment by two examiners of ten monthly compositions written by the pupils sampled (a total score from 20 markings). A second criterion, the assessment by teachers of each pupil's ability to write English (a five point scale) was also used but proved to be less helpful in that it correlated less well with the assessments of pupils obtained from the three experimental marking procedures. Despite this finding, the Schools Council reports a higher correlation between the teachers' assessments when scaled and the main criterion than between the official O-level marks of pupils awarded by the examining board and the main criterion; marginally higher correlations were obtained for the three experimental procedures (these latter differences might be attributable to the coarseness of the scale used to report the assessments by teachers). The finding that the assessments by teachers provided a higher 'validity' coefficient than the official examination results is by no means an isolated one. As seen earlier, evidence from studies of the entrance examination points to the usefulness of primary teachers' estimates as predictors of future attainment in secondary school. Also, a more recent study by Chedzoy (1976) provides strong evidence that the assessments by teachers of their pupils' A-level mathematics grades is a better predictor of success in university mathematics than the official A-level grades awarded

by an external examining body. As will be seen shortly, research in Australia and New Zealand shows that school assessments are at least as effective for predicting university success as results from external examinations.

The remaining discussion of British research concerns the finding that teachers tend more often to overestimate rather than underestimate their pupils' performance. This particular finding is either strongly supported, or supported in part, by three further studies, but is unsupported in one major study concerned with the monitoring of standards in examining at the CSE level. In the latter study (Schools Council 1966b) the estimated and achieved mean grades were reported for samples of pupils drawn from 9 CSE and 7 GCE examining boards. Overall, the results show a close agreement between the forecasts and actual grades awarded: of the six subjects studied in each sector (GCE and CSE) the maximum difference in mean grades (estimated and achieved) taking all boards together was 0.2 of a grade. A similar finding was obtained in another study in the case of three out of four examinations analysed (Bloomfield et al, 1977); in this investigation the grades awarded by teachers for Mode III CSE were related to the grades awarded by the examining board for Mode I CSE by means of a reference or monitoring test. In the fourth examination analysed, the gradings of teachers were found to be significantly more lenient than those of the examining board. In a study of A-level History, Ferguson and Garrett (1977) also report that teachers tend to overestimate their pupils' performance particularly for Grades A, B and C; no statistical details are presented but the finding that teachers awarded more A's than the examining board is perhaps surprising in view of the evidence reported earlier that teachers tend to underestimate the performance of pupils at the top end of the scale. Finally, in an experimental

study of assessment in O-level English (JMB, 1970) strong evidence is presented to show that teachers tend to leniency in grading their pupils. In this study the grades awarded from an externally administered examination paper were compared with those awarded by teachers; in all but one of nine schools sampled, a higher mean grade was awarded by the school than the examining board. In terms of the pass/fail criterion, the schools and examiners agreed in 332 out of 444 pupils sampled (74.8 per cent of cases); of the remaining 112 pupils, 108 were passed by the school but failed the examination (24.3 per cent of cases) whilst only 4 were failed by the school but passed the examination (0.9 per cent of cases). These figures are in fact more extreme than those reported at the pass/fail criterion by Petch.

Correlation and grade standards between schools

As seen, a number of researchers argue that teachers in general are able to rank their pupils, as measured by the correlation between the estimated and achieved marks of pupils in public examinations, but are unable to establish a standard for their pupils' work which places them in a position appropriate to other groups on a national standard. Murphy (1979) in fact questions whether teachers are able to produce a similar rank order to that of the examiners (admittedly the correlations he obtained were lower than those reported by Petch and McClelland). Setting aside this difference of opinion, the problems related to scaling differences are especially highlighted by a study carried out in Britain by Willmott and Hall (1975). This study compared the marks obtained by pupils in a variety of O-level examinations with the marks the pupils would have obtained had the markings of questions in the examinations been adjusted to provide questions of equal difficulty; (there are good reasons for such an adjustment where pupils are allowed a choice of questions). As such, the adjustments made little difference to the distributions (shape) of scores and even produced correlations

in every case in excess of 0.99 with the rank marks of pupils. However, these results masked a substantial variation between the two sets of marks in relation to the movement of candidates around the score values used to define grade boundaries. In the subjects analysed, the number of pupils who crossed the pass/fail boundary varied from 4.3 to 6.6 per cent of all candidates; on the full nine-point scale used to report results, the numbers changing by at least one grade ranged from 26.5 to 40.9 per cent of candidates. The message from these results is quite clear: if grade movements of this order are associated with correlations as high as 0.99, a good deal of caution is needed in drawing conclusions from correlations and values in the range 0.6 to 0.9 (i.e. values typically reported in the studies reviewed thus far). In other words, it is clear that an evaluation of teachers' assessments should incorporate more than a correlational analysis of the estimated and achieved marks of pupils; analyses of 'hits' and 'misses' at the pass/fail and other boundary points of the scale used is essential for a fuller picture to be obtained.

Research elsewhere overseas

Unfortunately, outside Britain much of the research carried out on the accuracy and validity of teachers' assessments has remained unpublished although not unrecorded. For example, Fairbairn et al (1976) reporting on the movement to internal assessment in Queensland, commented that ...

"Evidence available to the Radford Committee suggested that while individual schools would be capable of making accurate assessments of students, they would need assistance in making their standards comparable with those in other schools."

(p.143)

Similarly, Elley (1974 and 1976, respectively) reports:

"Recent analyses of fifth form teachers' assessments conducted at the NZCER show that they usually rank pupils in much the same order of ability as the external examinations. However, their standards of passes and grades are often found to vary from the national norm in an idiosyncratic way. These findings confirm overseas research which suggests that a system of moderation is needed if the assessments of different schools are to be comparable." (p.19)

"Many teachers still do not see the problem. They believe they can keep their assessments in line with those of other teachers. We have shown that this is not so in New Zealand - both at 5th and 6th form level.

The same experiment has been repeated in England, in Canada, in Australia. Repeatedly, teachers' assessments correlate well with external exam results on most subjects. It is just the standards that vary from school to school. Some are too lenient, some too severe. What teachers need is help with scaling." (p.29).

In the case of the Queensland evidence, a careful study of the Radford report (Radford, 1970) in fact fails to reveal figures or statistics which specifically relate teachers' assessments to concurrent or end-of-year examination results; references are made, however, to the need to scale teachers' assessments. Similarly, the New Zealand research referenced to by Elley has remained unpublished (and difficult to access*) whilst reports of the Australian and Canadian experience focus primarily on techniques for scaling or monitoring teachers'

* Personal communication between Dr C G W Hall (VUW) and the NZCER.

assessments rather than on the evidence which indicates that 'teachers can rank but not scale'. It should be added that the arguments made by Elley for scaling teachers' assessments draw in part upon personal contact and communication with the relevant state/provincial educational bodies in Australia, Canada and elsewhere.

Despite the lack of relevant published data, the argument ought to be made (on logical grounds anyway) that the widespread movement to internal assessment in several countries (see Elley, 1974, for example) implies general confidence and satisfaction with teachers' assessments: that is, that teachers are able to yield at least as valid an assessment of their pupils' worth as an external examination. As seen, the British experience supports this conclusion. Also, it cannot be argued that the validity of teachers' assessments is lessened by the need to incorporate scaling procedures; if teachers are in fact out of line with each other then the validity of their assessments must be supported by the safeguard of an appropriate scaling procedure. With these points made, it is relevant to look in brief at the results of the two studies for which 'hard' facts are available - the Radford Report showing the predictive (as distinct from concurrent) validity of teachers' assessments, and statistical evidence from Ontario pointing yet again to the relative leniency of the gradings made by teachers.

The Radford Report provides details of two analyses looking at the relative effectiveness of different predictors of success at University. In the first of these, success in first-year university mathematics was correlated with seventeen different measures including marks obtained in the Senior (State) Secondary Mathematics examination, marks obtained in a school based (constructed and conducted) examination, and teachers' estimates of their pupils' results in the Senior examination. The values obtained ranged from 0.62 to 0.15 with a median value of 0.49;

highest value was for the school based examination with comparable values being obtained for the other two measures (around 0.52 - 0.54). In the second study, performance in first year science and arts was correlated with aggregate grade in the Senior Examinations (best five subjects), aggregate mark in the same examinations, and an aggregate mark based on the predictions by teachers of their pupils' marks in the state examinations. Correlations of 0.43, 0.47 and 0.43 were obtained respectively between success in arts and the above predictions, and corresponding correlations of 0.42, 0.43 and 0.43 were obtained between success in sciences and these predictions. Whilst values in the range 0.43 to 0.47 (the range obtained here) cannot be considered high, the combined evidence of the marks and aggregate studies lends itself again to the conclusion that teachers' assessments are at least as valid as the marks/grades obtained from an externally conducted examination when it comes to the task of predicting future academic attainment.

In his review of overseas experiences with internal assessment, Elley (1976) reports evidence from Ontario that the percentages of pupils gaining UE increased substantially with the replacement of the Grade 13 (7th form equivalent) external examination by a system of teacher assessment without moderation. The actual figures are summarised in Table C7 of Watson and Quazi (1969) and show an increase of 5.7 per cent in the number of Grade Ones awarded, 8.6 per cent for the number of Grade Twos and 2.6 per cent for the number of Grade Threes - a 16.9 per cent increase overall in the number of passes as compared with the average of the previous five years examination results (the overall increase is slightly lower than the value of 20 per cent cited by Elley). Elley also reports that the percentage of students gaining the UE continued to rise in subsequent years but not without growing

concern from the University Admission Officers who pointed also to a substantial variation between schools in the decline of the standard of the school leaving certificates. Elley cites a similar finding in British Columbia where the University in fact introduced preliminary entrance tests in some faculties in order to compensate for the variation and decline in standards.

2.3 Research in New Zealand

Whilst the view that 'teachers can rank but not scale' is one which has been expressed in the New Zealand literature by a number of writers in a variety of publications - SCEB (1972, 1977), Tate (1977) and Lewis (1977) to name a few - it is evident that the major source of materials for this view are the previously cited publications by Elley and an earlier joint publication by Elley and Livingstone (1972) which outlined various alternative proposals for replacing external examinations by a system of internal assessment. It is unfortunate from the viewpoint of the present research that more statistical information was not provided by Elley in support of his arguments - it would appear that his data at the fifth and sixth form levels represent the most extensive coverage of schools and subjects in New Zealand to date. The New Zealand evidence which is to be dealt with here comprises:

1. three studies (two unpublished) which directly compare teachers' estimates with end-of-year examination results;
2. a study by Parkyn (1959) which investigated the predictive validity of teachers assessments, and
3. a study reported in Elley and Livingstone (1972) which emphasises the disparate marking standards of teachers.

The first New Zealand research identified by the writer which compares teachers' assessments with external examination results was carried out by the New Zealand Department of Education as part of an

evaluation of the 1969 School Certificate Science Examination (N.Z. Dept of Educ., 1971). The following extract summarises the relevant finding:

"Another measure of validity is provided by correlating the marks obtained from the examination with the internal assessments produced by schools before the examination. Using this approach it was found that the median correlation across several large schools was 0.84. In other words, the S.C. Examination in science produced results very similar to those obtained in local school examinations." (p.21)

Unfortunately no further details are provided, but it is interesting that the evaluators see the internal assessments of teachers as a criterion of validity!

The major New Zealand source material identified by the writer is an unpublished master's thesis by Hulbert (1978) concerned with the moderation of internal assessments for School Certificate. Hulbert sampled pupils from 10 schools (151 boys and 155 girls) and compared their performances in each subject taken for School Certificate with their rankings given by teachers (a nine point scale in each subject) and their results in final school examinations administered prior to School Certificate. The study also incorporated a number of standardised ability measures but in general these were found to correlate substantially less well with School Certificate results than either the teachers' rankings or the final school tests. Table 2.5 below summarises the correlations obtained by Hulbert between results in School Certificate and the two forms of assessment (final test and teacher ranking) in each of eight subjects analysed (subjects with a sufficient sample size). The 'Within' school correlations represent the median value for each subject from the schools sampled; the 'Between' schools values represent the rank order correlations between the paired mean scores

(assessed and achieved) for the ten schools sampled. This latter correlation in effect provides a measure of agreement between schools about the standards of their assessments; for example, a high correlation indicates that the order of schools as determined by the mean scores of their assessments of pupils agrees well with the order of schools as determined by the mean scores of their pupils in School Certificate. It should be recognised that this value focusses on the centre of the respective distributions and does not take account of variations in the standards at the extremes of the ability range. The remaining column of figures in Table 2.5 give the percentages for the number of disagreements between school assessments (final test) and School Certificate results in respect of pupils classified as pass or fail.

Table 2.5 Agreement between school assessments and School Certificate in eight subjects - from Hulbert (1978)*

Subject	School Certificate Marks with:				
	Final Test		Teachers' Ratings		Pass/fail discrepancies (%)
	Within	Between	Within	Between	
English	0.74	0.92	0.71	0.87	5
Mathematics	0.88	0.88	0.85	0.77	11
Science	0.80	0.95	0.72	0.93	19
Geography	0.85	0.32	0.80	0.62	16
Bookkeeping	0.81	0.17	0.82	0.16	16
Biology	0.88	0.20	0.87	-0.43	22
History	0.82	-0.20	0.76	0.67	22
French	0.72	0.86	0.76	0.68	20

An inspection of the values under the 'Within' columns shows overall a marginally higher set of values for the final test assessments than for the teachers' ratings. Ignoring this slight difference, it is apparent that the two sets of values provide a similar ordering of

* Taken from Tables 8.3 and 8.9.

subjects in respect of the size of the correlations obtained; Mathematics and Biology show closest agreement with School Certificate results whilst English and French show least agreement. On balance, the values obtained are of a similar order to that reported by Petch in England (see Table 2.1) although subject differences are quite marked in a few instances (compare the values for Biology and French in the two tables). The need to consider more than correlation coefficients is well illustrated by the discrepancy figures shown in the final column of Table 2.5. Values range from 5 per cent for English to 22 per cent for both Biology and History; the contrast between English and Biology is especially interesting in view of their relative correlations (0.74 and 0.88). This finding further reinforces the earlier argument concerning the need to use both correlation coefficients and discrepancy (expectancy) tables in order to obtain a clearer view of the agreement between school assessments and external examination results. It should be added that the discrepancy figures obtained by Hulbert supported only a slight bias towards leniency on the part of school assessments overall (see Tables 8.11 to 8.14 for further details).

Before the 'between' figures in Table 2.5 are considered it is interesting to compare the 'within' values obtained by Hulbert for English, Mathematics and Science with those obtained in an unpublished study by Gilmore* based on an analysis of assessments in five Dunedin high schools. The Gilmore study is of particular interest first, because of the size of the respective samples used - 38 classes in English, 37 in Mathematics and 29 in Science - and secondly, because like Hulbert she distinguished between assessments derived from scores in school examinations and teacher ratings of pupil performances in the form of

* Personal communication between Alison Gilmore (NZCER) and Dr C G W Hall (Educ. Dept V.U.W.)

a five-point grade scale (fewer classes were sampled for this second analysis). Unfortunately, Gilmore employed only the technique of correlation; consequently nothing can be said of the agreement between school assessments and School Certificate results at particular score points, e.g. the pass/fail boundary. The values she obtained are summarised in Table 2.6 - figures in parenthesis are the corresponding values reported by Hulbert.

Table 2.6 Summary of correlations obtained by Gilmore between school assessments and School Certificate marks

	Form of assessments	No. of classes	Correlation		
			highest	lowest	median
English	Exam	38	0.88	0.14	0.76 (0.74)
	Rating	11	0.84	0.53	0.77 (0.71)
Mathematics	Exam	37	0.99	0.20	0.83 (0.88)
	Rating	11	0.92	0.83	0.88 (0.85)
Science	Exam	29	0.94	0.08	0.73 (0.80)
	Rating	7	0.91	0.54	0.77 (0.72)

Two points emerge from Table 2.6: first, in contrast to Hulbert's results, Gilmore obtained marginally higher values for the ratings by teachers than for the assessments based on school examinations. This reversal, albeit a minor difference between the two studies, cannot be readily explained by the writer other than to point to differences between the rating instructions given to teachers - Gilmore actually instructed teachers to predict their pupils School Certificate results (graded A to E) whereas Hulbert asked teachers to compare each pupil with a 'representative' sample of 100 fifth form students in their first year. It should be added that there is no evidence elsewhere to indicate that variations in the instructions given affects the level of agreement between school assessments and external examination results (see the earlier discussion of Petch's study). Secondly, the range of

values obtained by Gilmore suggests that the examinations set by a few teachers (as distinct from the ratings given) were 'out of joint' with the examinations set by the School Certificate examiners. As argued earlier, a high correlation need not be expected in view of the different concerns of teachers and examiners, however, bearing in mind the general order of values given in Table 2.6, the minimum values of 0.20 (Mathematics), 0.14 (English) and 0.08 (Science) do not inspire confidence in the particular assessments concerned.

The 'between' schools correlations provided by Hulbert certainly confirm for some subjects anyway, that teachers differ in standards in making their assessments of pupils. Whilst few criticisms can be levelled at the final test correlations for English, Mathematics, Science and French, the values for Geography, Bookkeeping, Biology and History show little or no alignment between the assessed and achieved means of the different class groups. By and large the ratings of teachers fare little better; the values for Geography and especially History show a marked improvement upon the final test assessments, however the ratings for Biology show equally a deterioration in the between-school alignment of standards.

Further confirmation of the lack of agreement in standards comes from an earlier analysis of University Entrance assessments by Elley and Livingstone (1972). A similar technique to the 'between' schools analysis by Hulbert was used, but with the difference that comparisons were based on the percentage of students recommended for University Entrance in a given subject and the group's previous year's mean total (four best subjects) in School Certificate. Also, whereas Hulbert focussed on the variation in standards between schools in different subjects, Elley and Livingstone analysed the variation in standards between subjects in different schools (assessments in 12 schools were analysed). The correlations obtained by Elley and

Livingstone ranged from 0.90 to -0.51 with a median school value of 0.22; in summarising their results, Elley and Livingstone concluded:

"... without a common examination or moderation system, pupils who take mathematics, physical sciences and foreign languages will regularly be assessed more severely than they deserve."

(p.64)

Taking both the results of Hulbert and Elley and Livingstone together, the general conclusion to be reached is that standards of assessment vary not only between schools but also between subjects within the same school - a finding which echoes the research literature overseas.

The remaining study to be reported in this review is the 1959 investigation by Parkyn of success and failure at university. The study in question involved an analysis of the effectiveness of different variables in predicting university success at both first year and advanced year levels. The predictors included a number of standardised ability measures, aggregate mark in both School Certificate and University Entrance (for those who sat U.E.) and school assessments comprising sixth form rank (average of subjects taken for U.E.) and the principal's expectation of each pupil's success at university (a five-point scale). In general, highest correlations were obtained for the school assessments (the principal's assessments being marginally higher) and aggregate mark in School Certificate; not surprisingly, U.E. marks proved to be of little value because of the attenuated ability range of pupils in the analysis (only those pupils not accredited U.E. were included). The size of the correlations obtained for the school assessments ranged from around 0.35 to 0.40 (tan C*) with

* equivalent to Pearson Product-Moment Correlation Coefficient of around 0.50 - 0.55 (see Elley and Livingstone, p.70).

first year success (full-time students) to around 0.18 to 0.20 with advanced year success; lower values were obtained for part-time students. It is evident from these figures that none of the predictors proved to be particularly effective although the relative order of values obtained supports the findings reported earlier from overseas research (Radford in Australia and Chedzoy in England).

2.4 Summary

The preceding research may be summarised as follows:

1. Findings from both overseas and New Zealand show that the assessments of pupils by most teachers correlate well with the results of pupils in externally conducted examinations. In the main higher correlations have been reported for subjects with a more 'objective' criterion of assessments such as Mathematics (median correlations of around 0.85 to 0.90) than for subjects traditionally classified as part of the 'arts', e.g. English and History (median correlations of around 0.70 to 0.75).
2. Whilst the conclusion that 'teachers can rank their pupils' is reasonably justified by the research received, it should be recognised that a high correlation between teachers' assessments and examination results may well mask a substantial movement of pupils across key score/grade points in comparing the respective distributions. Also, a high correlation need not be expected between the two in order to conclude that a teacher is able to provide a valid ranking of his/her pupils; the different concerns of teachers and examiners need to be considered in evaluating the extent of agreement between distributions provided by each.
3. Expectancy table analyses of the level of agreement between teachers and examiners indicates that about 85 per cent of

pupils are classified to within one grade difference by the two (on a 5 or 6 point scale). In general, teachers tend more often to grade their pupils leniently, rather than severely, in comparing the grades awarded by examiners and teachers to the same pupils. Furthermore, substantial variations in standards of assessing are often found in comparing the assessments of any two (or more) teachers; the research points almost universally to the need for scaling or monitoring if a common standard of assessment amongst teachers is to be supported.

4. Lastly, evidence from a number of sources points quite strongly to the conclusion that teachers' assessments are at least the equal of external examinations when it comes to the task of predicting future educational achievement. This aspect of validity is clearly an important one in considering the case for the replacement of external examinations by a system (or systems) of internal assessment.

Chapter Three

DATA AND STATISTICAL ANALYSES

3.1 Introduction

The present chapter focusses on details on the data and statistical analyses used in this research. For the convenience of the reader the general aims of the study are repeated here:

1. to assess the degree of accuracy of teachers' predictions, in particular, at the pass/fail boundary, and
2. to identify and evaluate the relative importance of some correlates of accurate prediction.

Briefly, the first of these aims is investigated using the technique of correlation in conjunction with the construction of expectancy tables, and the second is investigated using step-wise multiple regression analysis (see Section 3.3).

3.2 Data

The data for this study are taken from the fifth form record book which shows for each class over the period 1973 to 1977 and subsequent to the commencement of this study, 1978.

1. the name of the teacher who took the class,
2. the mark in School Certificate for each pupil estimated by the teacher, and
3. the achieved mark of each pupil in School Certificate.

Other data collected for this study include:

Criterion variable

- V1 The Pearson Product-Moment Correlation Coefficient between the estimated and achieved marks (2 and 3 above) for each class.
- This value was chosen to represent a measure of the ability of a teacher to estimate accurately the performances of his/her

pupils in School Certificate.

Predictor variables

- V2 The sex of the teacher, classified as 1 (male) and 2 (female)
- V3 Length of teacher's service classified as:
 - 1: less than five years' service
 - 2: five to ten years' service
 - 3: greater than ten years' service
- V4 Commitment to subject by teacher. This variable identifies whether the subject is a first or second choice (major or minor) teaching subject for the teacher. It should be recognised that many teachers teach in subjects other than the one they have chosen to specialize in. This variable is classified as 1 (major) and 2 (minor).
- V5 Marked School Certificate. This variable identifies whether the teacher has marked School Certificate in the subject concerned and is classified as 1 (yes) and 2 (no).
- V6 Number of subjects presented for School Certificate, classified as 5 or 6. Classes comprise pupils enrolled in five only or six only School Certificate subjects; thus a class can be classified as a five-subject or a six-subject group.
- V7 Size of class, represented by the number of pupils.
- V8 Estimated ability level of the class; the mean of the estimated marks (percentage) for each class.
- V9 Achieved ability level of the class; the mean of the achieved marks in School Certificate for each class. Variables V8 and V9 were included to provide a measure of the overall ability level of each class - one measure being the teacher's mean estimate of the class's attainment (V8) and the other being the mean performance of the class in School Certificate. It was felt that the accuracy of a teacher's estimates of his/her pupils'

performance might vary with the overall ability level of the group concerned.

In choosing the above predictor variables for analysis it should be re-emphasized that the present study is, as far as the writer has been able to determine, the first in which an attempt has been made to formally identify characteristics associated with accuracy of prediction. The writer is therefore unable to hypothesize from previous research the importance or relative influence of the above variables; their selection in some cases represents no more than an inspirational guess as to a possible correlate (e.g. V2, V6, V8 and V9). In the case of V8 and V9 the writer has noted from his own experience that he is able to predict more reliably the results of more-able students than those from lower or average ability groups. In this context it should be noted that V6 (number of subjects presented for School Certificate) is associated with previous school attainment; in general, six-subject candidates have performed better in earlier school years than five-subject candidates. Other variables in the analysis (V3, V4, V5 and V7) offer a more logical basis for inclusion; one could hypothesize, if only tentatively, a correlation between accuracy of prediction and length of service (greater experience), commitment to the subject (major), experience in marking School Certificate, and the size of the class (smaller classes permit closer monitoring). One further point: data relating to the qualifications of teachers was also collected but it was found that all but a few teachers were trained qualified graduates, i.e. this variable afforded no basis for discrimination.

3.3 Statistical Analyses

The analyses for this research were carried out in four major steps:

1. The calculation of the Pearson Product-Moment Correlation

Coefficients between the estimated and achieved marks for each class (teacher) from 1973 to 1977. For this analysis four mainstream subjects were selected for study - English, Mathematics, Science and French. These subjects were chosen because they provide the major basis, in terms of subject intake, for determining future sixth form enrolments. Performance in School Certificate English, Mathematics and Science, separately or in combination, also form the core of a number of studies relating School Certificate to future academic attainment (see, for example, Parkyn (1959) and Marsh (1970) for studies of university success, and the earlier discussion of Elley and Livingstone (1972) concerning the monitoring of sixth-form assessments). In addition, it should be pointed out that the subjects chosen provide a relatively stable staffing situation within the school in that long-standing heads of department have been able to provide necessary details not included in the record book (such as years of experience of former members of staff).

2. The construction of expectancy tables relating the estimated marks of pupils to their achieved marks in each of the four subjects chosen for detailed study. These tables provide greater detail of the accuracy of teachers' estimates at different points across the estimated and achieved mark range, for example, the movement over the pass/fail boundary (see page 58 for an example). To ensure a sufficient sample size, data was pooled across teachers within each of the years 1974 to 1977 (for administrative purposes it was not possible to provide details for 1973). Thus a separate expectancy table is provided in each of the four subjects, English, Mathematics, Science and French, for each of the years 1974 to 1977.

In addition, expectancy tables have been drawn up relating the estimated number of subject passes to the achieved number of subject

passes of pupils (all subjects included) within each of the years 1974 to 1977 (see page 85 for an example).

3. The multiple regression analysis of the accuracy of teachers' estimates (V1) using variables V2 to V9 as predictors. The analysis was carried out separately for English, Mathematics, Science and French; data were combined across the years 1973 to 1977 to enable a sufficiently large sample of classes to be obtained (each class provides one measure of a teacher's accuracy in estimating pupils achievement). For the analysis, the forward stepwise regression option provided by the Statistical Package for the Social Sciences (data) was used. This procedure selects initially the predictor variable which explains the greatest amount of variance in the dependent variable; it then selects a second variable which in conjunction with the first explains the greatest amount of variance, and so on. In other words, at each step the variable that explains the greatest amount of residual variance is entered.

In the printout of the analysis for each subject, additional statistical information was requested and is reported later; for example, the intercorrelation matrix between all variables (V1 to V9 for each subject).

4. A summary analysis of data for 1978, that is, data which was not available at the time of the regression analysis described in Step 3 above. The correlations were obtained between the estimated and achieved marks of pupils in 14 subjects (a list of these is given in Table later) and compared with similarly obtained values for 1977.

Chapter Four

PRESENTATION OF RESULTS

4.1 Introduction

The major objective of this chapter is to present and describe the results of the analyses outlined in the previous chapter. In the main, the major interpretations and discussions are left to the next chapter. The results are presented in sections corresponding to the division of analyses described earlier, namely:

- 4.2 Correlation between the estimated and achieved marks of pupils in School Certificate for each class.
- 4.3 Expectancy table analysis of the estimated and achieved marks of pupils in School Certificate for each year.
- 4.4 Multiple regression analysis of the accuracy of teachers' estimates.
- 4.5 Summary analysis and comparison of data for 1977 and 1978.

4.2 Correlation between the estimated and achieved marks of pupils in School Certificate for each class

The results for this section are presented in Tables 4.1a to 4.5. Tables 4.1a to 4.1c show details for School Certificate English of:

- n number of pupils in the class,
- $\bar{x}_e$ mean score of the teacher's estimated marks of pupils in the class,
- $\bar{x}_a$ mean score of the achieved marks of pupils in the class,
- $\bar{x}_e - \bar{x}_a$ the difference between the means of the two sets of marks,
- r the Pearson Product-Moment Correlation Coefficient between the estimated and achieved marks of pupils in the class.

Table 4.1a shows details for 1973, Table 4.1b shows details for 1974, and so on. The same format is used for Mathematics (Tables 4.2a to 4.2e), Science (Tables 4.3a to 4.3e) and French (Tables 4.4a to 4.4e). The remaining table, Table 4.5, presents the mean and median correlations between the estimated and achieved scores of pupils summarized across classes for each subject and each year. It should be noted that the mean correlations are unweighted values.

Essentially there are two major pieces of information summarized in the tables, namely, the correlations and the differences between the estimated and achieved marks of pupils. The importance of the first is self-evident; correlations give directly a measure of each teacher's ability to estimate accurately the pupils' marks. Correlations alone, however, are not enough; a teacher may estimate accurately (even perfectly) the order of marks of pupils, but in each case may provide an over- or an underestimate of these marks. Where a teacher underestimates the marks of his/her pupils more students will actually pass School Certificate than indicated by the teacher's predictions; where a teacher overestimates the marks of his/her pupils the converse will occur. The difference between the mean of the estimated and achieved marks of pupils is therefore a useful piece of information to hold in conjunction with the correlations. The discussion will focus first on the correlations provided.

The two main points to emerge from the various tables of correlations are the greater degree of accuracy of predictions overall in Mathematics and French than in Science and English, and the greater consistency (i.e. less variation) in the predictions of teachers of Mathematics and French than teachers of Science and English. For example, the summary values given in Table 4.5 show a fairly consistent pattern across years of around 0.89 (Mathematics), 0.87 (French),

Table 4.1a Mean scores and correlations for English 1973

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	26	79.9	76.2	3.7	.60
2	28	71.9	66.3	5.6	.73
3	31	67.5	69.4	-1.9	.60
4	28	62.3	61.4	0.9	.55
5	30	58.8	62.8	-4.0	.72
6	30	56.1	53.8	2.3	.69
7	15	50.9	44.9	6.0	.66
8	8	50.0	47.1	2.9	.73
9	24	54.8	55.3	-0.5	.85
10	19	50.2	46.0	4.2	.90
11	22	46.4	43.5	2.9	.76
12	29	59.5	51.9	7.6	.68

Table 4.1b Mean scores and correlations for English 1974

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	30	76.2	74.5	1.7	.75
2	31	68.9	69.1	-0.2	.79
3	34	62.5	66.9	-4.4	.56
4	29	59.4	59.7	-0.3	.80
5	19	57.2	57.1	0.1	.72
6	21	58.9	60.5	-1.6	.69
7	20	51.8	55.6	-3.8	.81
8	23	59.4	60.8	-1.4	.54
9	24	51.5	45.7	5.8	.87
10	25	46.8	45.7	1.1	.76
11	23	46.5	52.7	-6.2	.85

Table 4.1c Mean scores and correlations for English 1975

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	30	76.3	75.9	0.4	0.72
2	20	69.6	67.6	2.0	.58
3	34	64.9	66.2	-1.3	.62
4	30	65.5	67.1	-1.6	.69
5	23	58.8	54.9	3.9	.50
6	19	54.3	52.9	1.4	.89
7	24	52.9	55.8	-2.9	.66
8	28	60.5	64.5	-6.0	.86
9	24	47.5	53.2	-5.7	.82
10	26	42.8	45.5	-2.7	.75
11	21	52.9	45.5	7.4	.75

Table 4.1d Mean scores and correlations for English 1976

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	33	65.0	69.3	-4.3	.85
2	27	67.3	66.9	0.4	.76
3	24	70.2	76.3	-6.1	.52
4	27	63.6	64.3	-0.7	.62
5	27	62.1	60.3	1.8	.89
6	30	55.7	52.5	3.2	.72
7	30	51.0	58.4	-7.4	.76
8	27	49.7	49.2	0.5	.59
9	18	48.1	44.6	3.5	.59
10	25	48.3	45.8	2.5	.76
11	24	49.3	40.5	9.3	.87

Table 4.1e Mean scores and correlations for English 1977

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	34	65.8	67.8	-2.5	.56
2	34	64.2	65.8	-1.6	.51
3	34	67.5	67.8	-0.3	.65
4	25	57.6	58.7	-3.9	.82
5	28	48.6	54.0	-5.4	.72
6	27	55.4	48.8	7.1	.71
7	27	51.8	50.8	1.0	.80
8	27	53.0	48.9	4.1	.85
9	29	53.0	51.2	1.8	.92
10	18	52.5	45.2	7.3	.41

Table 4.2a Mean scores and correlations for Mathematics 1973

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	25	67.4	79.2	-11.8	.94
2	26	66.7	67.8	-1.1	.88
3	31	66.5	78.4	-11.9	.94
4	25	44.4	47.1	-2.7	.93
5	27	53.0	56.0	-3.0	.87
6	25	46.6	43.5	3.1	.91
7	17	37.2	36.4	0.8	.71
8	23	48.4	43.2	5.2	.82
9	26	51.7	43.3	8.4	.88
10	9	51.1	43.3	7.8	.93

Table 4.2b Mean score and correlations for Mathematics 1974

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	29	70.8	77.0	-6.2	.86
2	31	68.3	69.1	-0.8	.89
3	33	57.0	63.6	-6.6	.85
4	31	56.6	49.9	6.7	.92
5	17	40.5	34.9	5.6	.74
6	21	47.9	50.2	-2.3	.89
7	20	52.3	46.6	5.7	.95
8	22	52.8	45.7	7.1	.82
9	23	34.8	37.9	-3.1	.91
10	16	40.4	32.8	7.6	.69

Table 4.2c Mean score and correlations for Mathematics 1975

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	31	60.3	69.7	-9.4	.67
2	27	70.3	75.8	-5.5	.95
3	32	57.1	64.6	-7.5	.83
4	31	68.1	68.5	-0.4	.88
5	20	44.0	45.9	-1.9	.93
6	19	38.2	48.2	-10.0	.76
7	24	47.6	50.6	-3.0	.89
8	28	48.5	55.3	-6.8	.90
9	24	40.6	39.0	1.6	.81
10	20	31.2	35.1	-3.9	.93

Table 4.2a Mean score and correlations for Mathematics 1976

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	24	58.6	71.8	-13.2	.89
2	32	63.2	75.8	-12.6	.91
3	28	68.8	62.3	6.5	.83
4	26	66.6	74.5	-7.9	.94
5	29	48.5	59.3	-10.8	.91
6	30	45.8	45.3	0.5	.94
7	31	38.4	51.0	-12.6	.80
8	26	37.7	41.7	-4.0	.87
9	16	29.1	39.6	-10.5	.93
10	14	47.7	36.6	11.1	.87

Table 4.2c Mean score and correlations for Mathematics 1977

Class	n	$\bar{x}_o$	$\bar{x}_a$	$\bar{x}_o - \bar{x}_a$	r
1	33	64.1	63.9	0.2	.93
2	33	58.7	66.6	-7.9	.87
3	34	62.4	59.5	2.9	.88
4	25	63.0	59.6	3.4	.95
5	28	45.4	45.4	0.0	.92
6	26	44.8	49.2	-4.4	.86
7	25	51.4	40.4	11.0	.94
8	20	51.4	42.7	8.7	.87
9	33	46.4	40.7	5.7	.56

Table 4.3a Mean scores and correlations for Science 1973

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	24	73.0	73.3	-0.3	.80
2	27	66.0	68.4	2.7	.88
3	30	62.5	66.2	-3.7	.71
4	18	56.7	51.4	5.3	.68
5	13	58.5	63.6	-5.1	.72
6	10	55.4	60.7	-5.3	.76
7	33	52.7	55.2	-2.7	.83
8	13	46.9	42.0	4.9	.41
9	21	53.0	53.8	-0.8	.66

Table 4.3b Mean scores and correlations for Science 1974

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	27	70.0	71.2	-1.2	.78
2	31	60.2	61.0	-0.8	.80
3	33	59.0	57.2	1.8	.68
4	31	53.2	51.6	1.6	.83
5	21	52.1	50.8	1.3	.61
6	20	53.2	48.2	5.0	.91
7	23	51.7	50.1	1.6	.90
8	30	47.0	43.1	3.9	.76

Table 4.3c Mean scores and correlations for Science 1975

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	30	64.0	68.8	-4.8	.83
2	28	62.3	66.9	-4.6	.82
3	34	57.4	62.4	-5.0	.71
4	30	56.3	61.5	-5.2	.83
5	27	51.4	51.2	0.2	.74
6	23	41.5	48.9	-7.4	.62
7	25	48.3	50.9	-2.6	.84
8	12	45.8	36.0	9.8	.76
9	15	49.1	40.3	8.8	.71

Table 4.3d Mean scores and correlations for Science 1976

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	24	63.1	70.5	-7.4	.75
2	33	63.8	59.6	4.2	.80
3	28	51.4	53.9	-2.5	.77
4	27	57.3	66.7	-9.4	.82
5	27	51.9	52.1	-0.3	.90
6	30	35.1	44.4	-9.3	.77
7	30	47.7	42.1	5.5	.76
8	26	47.0	38.7	8.3	.76
9	22	48.2	36.6	11.6	.80
10	16	42.5	31.5	11.0	.78

Table 4.3e Mean scores and correlations for Science 1977

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	32	68.1	66.9	-3.8	.84
2	34	56.2	58.7	-2.5	.72
3	33	60.2	62.9	-2.7	.91
4	25	56.0	50.0	6.0	.87
5	28	48.9	50.7	-1.8	.61
6	27	54.8	50.7	3.6	.72
7	25	41.0	48.1	-7.1	.81
8	22	46.1	44.5	1.6	.71
9	16	45.6	35.8	9.8	.89

Mean Scores and Correlations for French 1973-1977

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	27	63.2	61.9	1.3	.90
2	20	58.3	61.6	-3.3	.87
3	24	46.9	42.3	4.6	.88
4	25	42.4	38.2	4.2	.86
5	30	47.5	44.4	3.1	.76

4.4a 1973

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	29	61.4	72.7	-11.3	.87
2	16	56.1	56.4	-0.3	.92
3	31	48.9	46.7	2.2	.76
4	19	52.2	42.2	10.0	.84
5	10	45.0	36.7	8.7	.85

4.4b 1974

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	30	64.5	71.7	-7.2	.86
2	23	46.8	59.7	-12.9	.84
3	24	42.5	44.6	-2.1	.86
4	22	48.5	52.6	-4.1	.95
5	19	42.5	44.6	-2.1	.95

4.4c 1975

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	15	64.5	75.2	-10.7	.84
2	29	47.7	54.7	-7.0	.84
3	13	58.8	65.1	-6.3	.91
4	22	45.6	45.1	0.5	.91
5	29	43.1	43.6	-0.5	.89

4.4d 1976

Class	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
1	14	65.1	72.2	-7.1	.91
2	33	59.2	62.7	-3.5	.93
3	8	43.8	43.5	0.3	.84

4.4e 1977

Table 4.5

Mean and Median correlations
between the estimated and
achieved scores of pupils,
by subject for each year.

Year	English		Maths		Science		French	
	Mean*	Mdn	Mean*	Mdn	Mean*	Mdn	Mean*	Mdn
1973	.70	.71	.88	.90	.72	.76	.79	.87
1974	.74	.76	.85	.88	.78	.79	.85	.85
1975	.71	.72	.86	.89	.76	.76	.89	.86
1976	.72	.72	.89	.89	.79	.77	.88	.89
1977	.70	.72	.89	.90	.79	.81	.89	.91
Overall*	.71	.72	.89	.89	.77	.77	.86	.87

* Unweighted values

0.77 (Science) and 0.72 (English) for the mean and median correlations between the estimated and achieved scores of pupils. These clearly suggest a subject effect which will be discussed later.

The variation in the accuracy of teachers' predictions is most evident in English with values ranging from 0.60 to 0.90 for 1973 (the least variation) to 0.41 to 0.92 for 1977 (the most variation). Similarly obtained figures show values of 0.73 to 0.90 and 0.41 to 0.83 for Science, 0.80 to 0.94 and 0.56 to 0.95 for Mathematics, and 0.84 to 0.97 and 0.76 to 0.92 for French. The values presented show, in effect, a decrease in the variation from English to Science to Mathematics and finally to French. It should be pointed out, however, that the values for French were based on fewer classes than those for the other subjects and this possibly contributed to the greater homogeneity in the values obtained; this point is discussed further in the next chapter. A further point should be made in relation to the relatively atypical value of 0.56 obtained by one teacher of mathematics for the 1977 data; for the class concerned the estimate was made later in the year following shortly after a change of teacher. A look at Table 4.2e in fact shows that the next lowest value was 0.86! Unfortunately, changes of teachers have not been consistently recorded; therefore it is not possible to determine whether or not a systematic effect exists across subjects between change of teacher (number of changes and the timing of the change) and accuracy of prediction.

The results of the mean analysis, that is the comparison of the mean estimated and achieved scores of classes, reveals three major points of interest. First, it is evident that there is no clearly defined pattern within subjects between the ability of a teacher to order his/her pupils accurately (as measured by the correlations given) and the accuracy of the teacher in pinpointing his/her class's mean

mark in School Certificate. In some instances the most extreme case of over- or underestimating the class's performance is associated with a low correlation whilst in other instances the most extreme case is associated with a very high correlation. Secondly, a study of the magnitude of the differences between the mean estimated and achieved marks of pupils shows a possible difference across subjects, namely, that English teachers were generally more able to pinpoint their class's achieved mean with greater accuracy than French, Science and, in particular, Mathematics teachers. This particular result contrasts quite dramatically with the correlational findings reported earlier in this section. It should be noted that 40 per cent of the mean estimates fell within two marks (plus or minus) of the mean achieved marks of pupils in English, whereas corresponding percentages of 27, 22 and 16 were obtained respectively for Science, French and Mathematics. At the extreme, only 25 per cent of teachers' estimates exceeded a mean difference of five or more marks (plus or minus) in English compared with 39 per cent for French, 40 per cent for Science and 61 per cent for Mathematics. Furthermore, it should be noted that 20 per cent of the Mathematics estimates exceeded a ten mark mean difference compared with 17 per cent for French, four per cent for Science and zero per cent for English. In short, whilst the within subject values show no systematic pattern between the ability of teachers to predict accurately the order of their candidates and their ability to pinpoint their classes' mean ability (as measured by School Certificate), the values across subjects suggest a negative relationship between these factors - Mathematics teachers predicted with greatest accuracy the order of their pupils but with least accuracy the mean ability of their classes, whereas English teachers predicted with least accuracy the order of their pupils but with greatest accuracy the mean ability level of their classes. This finding is taken up for further discussion in

the following chapter.

The third point of interest from the mean difference analysis relates to the direction of the difference between the estimated and achieved mean marks of classes. In English, 58 per cent of the predicted mean class values were in fact overestimates of the obtained values in School Certificate: this compares with 47 per cent for Science, 43 per cent for Mathematics and 39 per cent for French. In other words, these results suggest a tendency (tendency rather than trend) for English teachers to overestimate the results of their pupils whereas, in particular, French and Mathematics teachers tend to underestimate the results of their pupils. It should be stressed that there are substantial variations both across years and across teachers (within subjects) to this pattern. It should also be added that the analysis relates to the mean values - a more systematic analysis across the full range of marks is given in the next section.

4.3 Expectancy table analysis of the estimated and achieved marks of pupils in School Certificate for each year.

The results for this section of the analysis are summarized in Tables 4.6a to Table 4.15. As can be seen from Table 4.6a, the data contained in a grouped frequency (expectancy) table relates the estimated scores of individuals to their achieved scores in terms of frequency counts for pairs of intervals; for example, in Table 4.6a (English, 1974), one student was given a predicted score of 19 or less and actually obtained such a score; another student was also predicted to score 19 or less but actually achieved a score in the range 20-29. Similarly, out of three students with predicted scores of 20-29, two actually obtained such a score whilst the third scored in the interval 30-39. With the exception of the top and bottom intervals (80+ and 0-19) the intervals chosen divide the score range into ten-point groupings.

The diagonal entries (from bottom left to top right) represent students correctly classified in terms of the match between their estimated and achieved scores (i.e. correctly classified in terms of the intervals chosen). Entries in cells above (to the left) of the diagonal represents the counts of students whose predicted scores were overestimated relative to their achieved scores; entries in cells below (to the right) of the diagonal represent the counts of students whose predicted scores were underestimated relative to their achieved scores. The double lines in the table represent the cut-off point for the pass-fail boundary - a point of special interest to this study. At the foot of Table 4.6a figures are given for the total group size ($n = 279$), for the correlation between the estimated and achieved scores of students ($r = 0.78$) based on the interval classification used (i.e. not on the raw scores of students), and the percentage of students who crossed the pass/fail boundary line with respect to differences in their estimated and achieved scores (14 per cent). Tables 4.6b, 4.6c and 4.6d show the results similarly obtained for English in 1975, 1976 and 1977 respectively. Similar information is presented for Mathematics (Tables 4.7a to 4.7d), Science (Tables 4.8a to 4.8d) and French (Tables 4.9a to 4.9d). A summary table of the movement of subjects across the pass/fail boundary is contained in Table 4.10. This shows for each of the four subjects in each year the percentage of students predicted to pass but who failed (Pf), the percentage of students predicted to fail but who passed (Fp), and the total movement over the pass/fail boundary ($Pf + Fp$). The remaining tables and figures in this section will be described later.

Again there appear to be three major points of interest in looking at the tables mentioned. First, the correlations based on the interval classification of students provide a similar between-subject pattern to that described earlier for the median class correlat-

Table 4.6a: Grouped frequency (expectancy) table relating estimated mark to achieved mark, English 1974

Achieved %

	-19	29	39	49	59	69	79	80+
80						1	18	10
79					2	17	23	5
69				5	10	35	22	2
59			4	13	26	31	4	
49	1	2	14	5	11	5		
39	1	2	5	5				
29		2	1					
19	1	1						

Estimated %

n = 279 r = 0.78
 14% crossed pass/fail boundary

Table 4.6b : Grouped frequency (expectancy)
table relating estimated mark to
achieved mark English 1975

Achieved %

	-19	29	39	49	59	69	79	80+
80						1	7	11
79				1	4	14	37	8
69			1	3	24	25	17	4
59			9	17	23	22	4	
49		2	5	14	9	2	1	
39		4	9	3	2			
29	1	2	1					
19								

Estimated %

$n = 287$ $r = 0.77$
16% crossed pass/fail boundary

Table 4.6c: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, English 1976

		Achieved %							
		- 19	29	39	49	59	69	79	80+
Estimated %	80						1	9	11
	79					1	8	19	10
	69			2	3	17	37	17	5
	59			12	24	34	20	6	
	49		3	16	14	11	1		
	39		2	9	2	3			
	29		1	1					
	19		1						

$n = 300$
19% crossed

$r = 0.78$
pass/fail boundary

Table 4.6d: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, English 1977

Achieved %

Estimated %		19	29	39	49	59	69	79	80+
	80						1	4	7
	79				1	2	12	24	2
	69			1	6	28	35	18	3
	59			7	23	28	16	3	
	49		3	22	11	13	1		
	39		2	6	1				
	29		2	1					
	19	1							

$n = 284$

$r = 0.78$

18% crossed pass/fail boundary

Table 4.7a: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Mathematics 1974

		Achieved %							
		-19	29	39	49	59	69	79	80+
Estimated %	80						1	5	21
	79						5	14	11
	69				5	6	15	13	6
	59		3	8	12	14	10	7	
	49		11	11	10	3	1		
	39	3	7	7	3	1	1		
	29	7	10	3	1				
	19	6	3	1					

$n = 245$ $r = 0.87$
14% crossed pass/fail boundary

Table 4.7b: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Mathematics 1975

		Achieved %							
Estimated %		-19	29	39	49	59	69	79	80+
	80							8	21
	79						1	14	14
	69			1	1	2	22	14	4
	59			2	10	16	12	9	
	49		1	6	13	11	9		
	39	2	3	9	6	5	2		1
	29	3	7	2		1			
	19	15	7	1	1				

$n = 256$ $r = 0.88$
 17% crossed pass/fail boundary

Table 4.7c : Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Mathematics 1976

Achieved %

	-19	29	39	49	59	69	79	80+
80						2	3	22
79					4	3	8	16
69			1	1	5	8	12	7
59				4	20	16	11	3
49		4	9	11	19	6	1	
39	1	6	9	3	7			
29	5	7	6	2	1			
19	5	3	2			1		

$n = 257$

$r = 0.84$

16% crossed pass/fail boundary

Table 4.7d: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Mathematics 1977

		Achieved %							
Estimated %		-19	29	39	49	59	69	79	80+
	80						2	9	20
	79					2	9	21	5
	69			1	3	15	15	6	2
	59	1	2	6	10	22	7	4	
	49	1	7	15	16	7	2		
	39	1	15	10	7	1			
	29	4	2	3	1				
	19	3	2						

$n = 259$ $r = 0.85$
 13% crossed pass/fail boundary

Table 4.8a: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Science 1974

		Achieved %							
		-19	29	39	49	59	69	79	80+
Estimated %	80							4	5
	79					2	10	16	3
	69			2	2	13	21	12	1
	59			5	17	26	18		
	49		6	18	9	5	2		
	39	2	9	5	3		1		
	29								
	19								

$n = 217$ $r = 0.81$
 16% crossed pass/fail boundary

Table 4.8b : Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Science 1975

Achieved %

	-19	29	39	49	59	69	79	80+
80							3	10
79						7	16	5
69				1	7	24	13	4
59			4	11	32	26	8	
49		4	19	13	6	6		
39		8	8	5	3			
29		2		1				
19								

Estimated %

$n = 246$ $r = 0.82$
13% crossed pass/fail boundary

Table 4.8c: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Science 1976

		Achieved %							
		-19	29	39	49	59	69	79	80+
Estimated %	80							2	11
	79					2	4	10	6
	69			2		12	7	11	2
	59		2	9	19	15	17	8	1
	49	3	16	29	15	10	3		
	39		12	12	10	2	2		
	29		4	3		2			
	19		1						

$n = 264$ $r = 0.75$
 19% crossed pass/fail boundary

Table 4.8d: Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, Science 1977

		Achieved %							
		- 19	29	39	49	59	69	79	80+
Estimated %	80							1	11
	79				1		5	15	4
	69			1	2	11	9	14	3
	59			13	21	22	26	6	
	49		6	19	11	15	4		
	39	3	4	12	3	2			
	29		1						
	19								

$n = 245$

$r = 0.77$

24% crossed pass/fail boundary

Table 4.9a Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, French 19/4
Achieved %

	-19	29	39	49	59	69	79	80+
80							1	5
79				1	1	5	2	6
69				2	4	2	7	2
59			4	7	8	5	5	
49		4	10	11	5	8		
39		4	6	1	3			
29	1	1	4	3				
19	4		1					

$n = 133$ $r = 0.79$
23% crossed pass/fail boundary

Table 4.9b Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, French 1975
Achieved %

	-19	29	39	49	59	69	79	80+
80							2	6
79							1	5
69				1	1	5	9	2
59				2	12	11	3	2
49			3	7	9	2		
39		6	8	7	3			
29	1	3	2	2				
19		1						

$n = 116$

$r = 0.86$

15% crossed pass/fail boundary

Table 4.9c Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, French 1976
Achieved %

	-19	29	39	49	59	69	79	80+
80							2	5
79						1	5	3
69					4	6	5	1
59			2	4	13	5	2	
49			6	9	7	1		
39		1	7	3	3	1		
29	2	5		3				
19	1	1						

$n = 108$ $r = 0.85$
17% crossed pass/fail boundary

Table 4.9a Grouped frequency (expectancy)
table relating estimated mark to
achieved mark, French 1977
Achieved %

	-19	29	39	49	59	69	79	80+
80							4	6
79						3	5	2
69						3	5	
59				1	7	3	1	
49				3	3	1		
39			2	3				
29	1		1	2				
19		1						

$n = 57$ $r = 0.90$
9% crossed pass/fail boundary

t over

French		
P _f	F _p	T
11	12	23
3	12	15
6	11	17
2	7	9

it failed
it passed
boundary

ions (see Table 4.5). Overall, higher values were obtained for Mathematics (0.84 to 0.88) and French (0.79 to 0.90) than for English (0.77 to 0.78) and Science (0.75 to 0.81). Relative to the median values in Table 4.5, these correlations show a slight increase for English and Science and a slight decrease for Mathematics and French.

The second point to note is that the above variation in correlation between subjects is associated with at best only a slight difference in the percentage of students who crossed the pass/fail boundary. For example, over the four year period covered by the tabulated data approximately 15 per cent of Mathematics students, 16 per cent of French students, 17 per cent of Science students and 18 per cent of English students crossed the pass/fail boundary. However, these values mark an interesting variation between subjects in respect of the pattern of 'hits and misses' at the boundary point. In both English and Science approximately twice as many students were predicted to pass but failed as students who were predicted to fail but passed. This particular pattern was observed across years although not always with the same strength. In French almost the reverse pattern was observed; approximately twice as many students were predicted to fail but passed as students who were predicted to pass but failed. Again a variation in the magnitude of this pattern was observed across years. In Mathematics a less consistent result was obtained; in both 1974 and 1977 substantially greater numbers of students were predicted to pass but failed than the reverse, whilst in 1975 and 1976, substantially greater numbers of students were predicted to fail but passed than the reverse. Further discussion of these findings, in particular the variation in the results for Mathematics, is given in the next chapter.

The remaining point of discussion relates to the degree of over-

or underestimation of students' scores across the full range of marks in each subject. As mentioned previously, the cell entries to the left of (or above) the diagonal represent overestimates of performance whilst the cell entries to the right of (or below) the diagonal represent underestimates of performance. In respect of the entries provided, in both English and Science a greater number of overestimates than underestimates were given; in French and Mathematics a greater number of underestimates than overestimates were given. The results of this analysis are summarized in Table 4.10b showing for each subject the percentages of the total number of cell entries corresponding to under- and overestimates of performance. It should be noted that the results obtained showed a consistent pattern across years for French and only minor variations for English and Science. Again in Mathematics, results for 1974 and 1977 indicated substantially greater numbers of over- than underestimates, whilst for 1975 and 1976 the results show substantially greater numbers of under- than overestimates. Another feature brought out by the expectancy table analysis is that, in comparing the number of underestimates and overestimates, there was a greater tendency for teachers in English, Mathematics and Science to underestimate the performances of the more-able students and overestimate the performances of the less-able students. In French a more uniform pattern prevailed across the mark range, that is, proportionately larger numbers of underestimates than overestimates were given. Further evidence of this pattern for subjects is illustrated in the figures showing the histograms and cumulative frequency distributions for the estimated and achieved performances of students (Figures 4.1a to 4.5b). The histograms for the achieved scores of students (all years combined) in English, Mathematics and Science show greater dispersion, indicated by a smaller central peak and slightly larger areas towards the extremes, than the histograms for the estimated scores of students. In

Table 4.10b :

Summary of percentages
of pupils over and
underestimated 1974-1977

	Eng	Maths	Science	French
Percentage overestimated	34	27	36	22
Percentage underestimated	26	32	28	40

Figure 4.1a Frequency histogram for English for the years 1974 to 1977 combined

Figure 4.1b Cumulative frequency distribution for English for the years 1974 to 1977 combined

Figure 4.2a Frequency histogram for Mathematics for the years 1974 to 1977 combined

Figure 4.2b Cumulative frequency distribution for Mathematics for the years 1974 to 1977 combined

Figure 4.3a Frequency histogram for Science for the years 1974 to 1977 combined

Figure 4.3b Cumulative frequency distribution for Science for the years 1974 to 1977 combined

Figure 4.4a Frequency histogram for
French for the years 1974
to 1977 combined

Figure 4.4b Cumulative frequency
distribution for French for the years
1974 to 1977 combined

Figure 4.5a Frequency histogram for
Eng, Math, Sc, Fr for the years 1974
to 1977 combined

Figure 4.5b Cumulative frequency
distribution for En, Ma, Sc, Fr for the years
1974 to 1977 combined

French the histograms show a slight increase in the area above the centre ($> 60\%$) associated with a slight decrease in the area below the centre ($< 50\%$) in moving from the histogram for the estimated scores to the histogram for the achieved scores. These same effects can be seen in the cumulative frequency distributions - in English, Science and Mathematics the curves show a clear pattern of over-estimation at the lower end (the dotted line is to the right) and underestimation at the top end (the dotted line is to the left). In French, the dotted line appears to the left for much of the graph. Figures 4.5a and 4.5b show the combined data for the four subjects over the four year period. It should be noted that this figure represents the results of comparing nearly four and a half thousand estimated and achieved pairs of scores; it is evident that the pattern for English, Mathematics and Science dominates the graphs (there were substantially more entries in these subjects than in French).

The results discussed so far in this section have concentrated on the predictions of teachers within the four mainstream subjects of English, Mathematics, Science and French. Attention now turns to an analysis of the estimated number of subject passes in relation to the achieved number of subject passes based on the predictions of teachers in thirteen subjects - a list of these subjects is given in Table 4.21 but excludes Music. The major objective of this section of the analysis was to investigate the accuracy of teachers' predictions in the context of the criterion used for admitting pupils to the Sixth Form. For the purposes of the present study, entry to the Sixth Form is defined by a pass in three or more subjects, although it should be acknowledged that in practice a more flexible approach is often made (for example, some pupils may concurrently study at both the Sixth and Fifth form levels). The point should be made that pupils with a pass in three subjects are virtually assured of a place

in the Sixth Form, whilst students with fewer than three passes are evaluated on the basis of individual merit at the principal's discretion.

The format of the data is much the same as used in Tables 4.6a to 4.9d. In each table (Table 4.11 to 4.14) the estimated number of subject passes are related to the achieved number of subject passes in the form of an expectancy table. Thus, for example, for 1974 (Table 4.11) 23 pupils were estimated to pass in no subjects and achieved this end! A further nine students achieved one pass, another student two passes, having all been estimated to pass no subjects. The table also gives the number of pupils in the analysis ($n = 285$), the correlation between the estimated and achieved number of subject passes for pupils, ($r = 0.90$), and the percentage of students crossing the boundary line for entry to the Sixth Form (11 per cent). One further table of relevance is Table 4.15 which summarizes the percentages of students crossing Sixth Form entry criterion for the four years under study.

The first point to note from the Tables for each year is that the correlation between the estimated and achieved number of subject passes ranged from 0.90 (1974) to 0.85 (1977). Clearly there are some extreme cases of failure to predict a pupil's ultimate performance (shown for example in the row corresponding to an estimated pass in one subject for 1977), however, it is apparent that the resultant correlations are heavily weighted by the large numbers for the cell entries at the extremes of the main diagonal. In other words, it would appear that teachers are more able to predict the results for the more-able and the less-able pupils than for the 'average' pupils - this is not really surprising.

A second feature of the results is the relatively stable value for the percentage of pupils crossing the Sixth Form entry criterion,

Table 4.11 relating the estimated number of subject passes to the achieved number of subject passes in 1974

Achieved no. subject passes

	0	1	2	3	4	5	6
6					2	5	35
5			1	4	14	40	13
4		1	4	19	19	5	2
3		3	11	8	3	3	
2	1	6	12	5	3		
1	11	11	7	3	1		
0	23	9	1				

$n = 285$

$r = 0.90$

11% crossed 6th form entry boundary

Table 4.12 relating the estimated number of subject passes to the achieved number of subject passes in 1975

Achieved no. subject passes

	0	1	2	3	4	5	6
6					2	10	39
5				1	9	44	14
4		1	4	9	12	15	1
3	1	6	10	9	10	6	
2	10	14	8	9	2		
1	8	7	5	3	1		
0	21	3					

$n = 294$

$r = 0.88$

13% crossed 6th form entry boundary

Table 4.13 relating the estimated
number of subject passes to the
achieved number of subject
passes in 1976

Achieved no. subject passes

Estimated no. subject passes	0	1	2	3	4	5	6
6					2	9	35
5			1	3	13	32	6
4			1	6	20	15	2
3	3	5	9	14	14	2	
2	4	9	11	11	7	1	
1	10	9	11	3	1		
0	30	7	1	1			

$n = 308$

$r = 0.88$

14% crossed 6th form entry
boundary

Table 4.14 relating the estimated number of subject passes to the achieved number of subject passes in 1977

		Achieved no. subject passes						
Estimated no. subject passes		0	1	2	3	4	5	6
	6					1	8	44
	5			1	11	14	35	13
	4	1	4	3	11	10	8	
	3	1	8	9	4	8	1	
	2	5	8	9	5	2	1	
	1	14	16	8	2	1	1	1
	0	16	4	3				

$n = 291$

$r = 0.85$

14% crossed 6th form entry boundary

Table 4.15
Summary of movement
over 6th form entry
boundary

Year	P_e	F_p	T	n
1974	7	4	11	285
1975	8	6	13	294
1976	6	8	14	308
1977	9	5	14	291
Total	8	5	13	1177

P_e Percentage of pupils estimated to qualify for 6th form entry but failed to do so.

F_p Percentage of pupils estimated to fail 6th form entry but passed

T Total percentage movement over pass/fail boundary

namely around 13-14 per cent. The results also show (see Table 4.15) a tendency overall for teachers to overestimate the number of subject passes although a variation on this pattern was observed for the 1976 data. This pattern is supported by a further analysis of the relative size of the cell entries above and below the main diagonal (i.e. not just in terms of the numbers of students crossing the Sixth Form entry criterion). Overall, the number of subject passes were overestimated for 29 per cent of pupils and underestimated for 22 per cent of pupils; the number of subject passes were correctly estimated for 49 per cent of pupils. Again, the general pattern of overestimating the performances of students was not observed in the 1976 data (here the number of subject passes were underestimated for 24 per cent of pupils and overestimated for 27 per cent of pupils). In general, the results suggest a tendency for the performances of the less-able students to be overestimated (most clearly observable in the data for 1975); however, no discernible pattern across years could be derived in respect of the performances of the more-able students.

The results in Tables 4.11 to 4.15 may be summarized then as showing a relatively high correlation between the estimated and achieved number of subject passes of pupils (around 0.88) associated with a 13-14 per cent shift in students crossing the Sixth Form entry criterion. Not surprisingly, more accurate predictions were made for the more-able and less-able students; however, there was a general tendency for more over- than underestimates of performance to be made, in particular, at the lower and middle ranges of the scale used (i.e. the scale of 0-6 subject passes).

4.4 Multiple regression analysis of the accuracy of teachers' estimates

For this section of the study the results are summarized in Tables 4.16 to 4.20. Tables 4.16 to 4.19 provide for each subject in

turn the intercorrelation matrix for the criterion variable (the correlation between the estimated marks of pupils in a class and their achieved marks) and the eight predictor (independent) variables. In studying the values shown, care should be taken to ensure that the direction (sign) of the various correlations is interpreted correctly. Table 4.20 provides details of the step-wise regression analysis of the ability of teachers to estimate their pupils' achieved scores in School Certificate. This Table summarizes for each subject:

1. The number of pupils or individual estimates,
2. the number of classes, that is, the n for the analysis,
3. the mean correlation between the estimated and achieved marks of pupils across all classes,
4. the standard deviation of the individual class correlations,
5. the multiple regression (predictor) coefficient from all variables,
6. the multiple regression coefficient squared, that is, the proportion of shared variance, and
7. the order of presentation of variables step-wise and their shared cumulative variance with the criterion variable.

Each Table will be discussed in turn.

The correlation matrix for English shows by and large a low order of values between the individual predictor variables and the criterion under study. Only the three variables V6, number of subjects taken, V8, the estimated mean of the class, and V9, the achieved mean of the class, reach a value of 0.30 (in fact all correlated negatively). It should be remembered that these three variables each provide an approximate measure of the ability overall of a class. This is reinforced in the matrix by their high correlations with each other, namely, 0.83 between V6 and V8, 0.84 between V6

Table 4.16
 * Correlation matrix for
 English 1973 - 1977 combined

	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
V_1 Corcoef	-04	-01	10	01	-34	-09	-31	-31
V_2 Sex		00	32	-15	-06	-07	07	-02
V_3 Service			-23	-50	-02	04	04	01
V_4 Commit				07	-14	11	-16	-08
V_5 MarkedSC					-03	-18	-18	-22
V_6 5,6 Subj.						51	83	84
V_7 No in class							46	64
V_8 Est Mean								91
V_9 Ach Mean								

* decimal points omitted

Significance levels for:

$$P = 0.05 \quad r = 0.27$$

$$P = 0.01 \quad r = 0.35$$

and V9, and 0.91 between V8 and V9. The only other correlations in the matrix to reach a significance of $p = 0.05$ (ignoring sign) are those between V2 and V4 ($r = 0.32$ - sex and commitment to the subject), V3 and V5 ($r = -0.50$ - teacher's service or experience and marking experience in School Certificate), and V7 (class size) with V6 ($r = 0.51$), V8 ($r = 0.46$) and V9 ($r = 0.54$). The first of these correlations (V2 and V4) suggests to a limited extent a greater proportion of male teachers with a major, and female teachers with a minor, commitment to the subject. The second correlation (V3 and V5) indicates a positive (N.B. not negative) relationship between length of teaching service and experience in marking School Certificate - neither of these variables individually correlate with the ability of teachers to predict their pupils' marks in School Certificate English. The remaining relationships link size of class to the ability of the class (V7 with V6, V8 and V9); it should be noted that in many subjects within the school there is an unwritten policy to place fewer pupils in classes with a preponderance of less-able students - these correlations reflect this situation.

As in the case for English, the correlation matrix for Mathematics generally shows a low order of values between the predictor variables and the criterion under study. Only the correlation between V4 (commitment to the subject) and the criterion reaches significance at the 0.05 level (-0.35) although the values for V8 and V9 with the criterion approach this level. The negative correlation between V4 and V1 in fact indicates a positive relationship between commitment to the subject and the ability of teachers to predict their students' results. The only other correlations of substance link sex with both commitment to the subject (greater commitment for females) and marking experience in School Certificate (males having had greater experience), and both length of service (V3) and class size (V7) with the trio of

Table 4.17

* Correlation matrix for
Mathematics 1973 - 1977 combined

	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
v_1 Cor coef	21	14	-35	-08	05	01	25	26
v_2 Sex		08	-45	40	-18	09	-12	-07
v_3 Service			-01	-26	66	45	63	64
v_4 Commit				28	24	-09	06	17
v_5 Marked SC					01	-08	-19	-01
v_6 5,6 Subj.						47	68	79
v_7 No in class							52	61
v_8 Est Mean								86
v_9 Ach Mean								

* decimal points omitted

Significance levels for

$$P = 0.05 \quad r = 0.28$$

$$P = 0.01 \quad r = 0.37$$

ability measures (V6, V8 and V9). The highest correlations in Table 4.17 are in fact between these three variables, namely, 0.68 between V6 and V8, 0.79 between V6 and V9, and 0.86 between V8 and V9. Some of these correlations clearly pose a few interesting questions - why should sex be related to subject commitment? (One might argue on traditional grounds that the relationship should have been in the reverse - Mathematics having been traditionally viewed as a 'man's domain'!) Also, what interpretations should be placed on the link between the teacher's experience and the ability of pupils? It should be noted that no such relationship was found in English although a moderate link exists in both Science and French (see below). Further discussion follows in the next chapter.

The results for Science show a number of significant relationships between the predictor variables and the criterion. These include: the three ability measures, V6 ($r = 0.70$), V8 ($r = 0.64$) and V9 ($r = 0.54$); class size ($r = 0.55$); commitment to the subject ($r = 0.56$, N.B. this is in fact a negative link - less commitment being related to better predictions by teachers of their pupils' achievements); length of service ($r = 0.44$); and lastly, marking experience in School Certificate ($r = 0.32$, again this is a negative link). Outside of these correlations, there are again strong links between the three ability measures, between both length of service and class size with these measures, and between (albeit moderately) both sex and commitment to the subject and these measures. In effect, these latter correlations suggest that female teachers, more so than male teachers, are more closely associated with 'able' groups, and similarly that the 'less committed' teachers, more so than the 'more committed' teachers, are more closely associated with 'able' groups. As before, it is not intended to offer any interpretations of these

Table 4.18
 * Correlation matrix for
 Science 1973 - 1977 combined

	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
v_1 Corcoef	27	44	56	32	70	55	64	54
v_2 Sex		25	22	46	42	14	42	33
v_3 Service			25	14	41	30	38	40
v_4 Commit				39	58	20	41	25
v_5 MarkedSC					32	00	20	00
v_6 5,6 Subj.						62	86	78
v_7 No in class							51	55
v_8 Est Mean								90
v_9 Ach Mean								

* decimal points omitted

Significance levels for:

$P = 0.05$ $r = 0.30$

$P = 0.01$ $r = 0.38$

results at this point although clearly there are some interesting questions posed.

The remaining table of correlations, for French, show roughly the same level of correlation as that found in the matrices for English and Mathematics. The only correlation of substance with the criterion variable is that for experience in marking School Certificate ($r = 0.42$) - it should be noted that the value indicates that teachers with no experience estimate their pupils' results better than those teachers with some experience! Again the table shows a link between the three ability measures and between length of service and these measures; other significant pairings include sex and length of teaching experience (males having had longer teaching experience), length of service and subject commitment (N.B. a positive, not negative, relationship), commitment to the subject and number of subjects taken by pupils for School Certificate (again a positive rather than a negative relationship), and experience in marking School Certificate and the estimated mean scores of classes (higher means being associated with no experience in marking at School Certificate). Again some of these results pose interesting questions although not all the results lend themselves immediately to meaningful interpretations (see next chapter). One final point should be made concerning the results for French - the various values are based on a total of only 23 classes (i.e. $N = 23$) and should therefore be treated with caution.

The discussion now turns to the contents of Table 4.20, namely, the step-wise regression analysis for each subject. Before discussing the main results, that is, the value of the multiple r and the order of presentation of variables for each analysis, two minor points of interest warrant brief attention. The first of these, the values under $\bar{r}$ (the mean correlation between the estimated and achieved marks of pupils across all classes) provides further reinforcement for

Table 4.19
 * Correlation matrix for
 French 1973 - 1977 combined

	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
v_1 Corcoef	22	05	-17	42	02	-20	15	20
v_2 Sex		-51	18	37	-02	04	32	28
v_3 Service			-50	-04	50	09	34	39
v_4 Commit				32	-42	-34	-13	-38
v_5 MarkedSC					-08	-30	55	36
v_6 5,6 Subj.						30	55	64
v_7 No in class							01	06
v_8 Est Mean								90
v_9 Ach Mean								

* decimal points omitted

Significance levels for

$$P = 0.05 \quad r = 0.41$$

$$P = 0.01 \quad r = 0.53$$

the conclusion that more accurate predictions were made in Mathematics and French than Science and English. Secondly, the greater variation in the accuracy of the predictions in Science and English is further evidenced by the standard deviations given in Table 4.20. Of particular interest here is the apparent lack of a systematic association between the standard deviations and the values for the multiple correlations - one might expect, although it does not necessarily follow, that a greater variation in the accuracy of predictions of teachers (as in English and Science) would provide more opportunity for correlation of the dependent with the independent variables. This does not, in fact, occur; the multiple correlation is greatest for Science and least for English.

To some extent the multiple correlations reflect the results in Tables 4.16 to 4.19. As seen, substantially larger correlations between the dependent and independent variables were obtained in Science than in the other subjects under study, although in the case of Mathematics and French, the multiple correlations suggest a masking effect amongst the variables. It is impossible from the data available to identify precisely the mechanisms involved in this effect, but clearly the multiple correlations for both Mathematics and French are higher than what one might expect from the correlations given in their respective matrices.

With respect to the order of presentation of variables in the step-wise analysis, it is difficult to discern a systematic effect across subjects although two variables appear more influential overall than the rest, namely, number of subjects taken for School Certificate (V6) and teachers' commitment to the subject (V4). If the first of these is seen as one of three measures of pupil ability in the analysis (bearing in mind the high correlations generally

Table 4.20 Multiple regression analysis of the accuracy of teachers' estimates of pupils' marks by subject, 1973 to 1977 combined

Subject	no. of pupils	no. of classes n	$\bar{r}_{v_i}$	s.d. v_i	mult r	% variance	Order of variables presented step-wise and % variance															
Eng.	1452	55	.72	.12	.37	14	V6	11	V7	12	V8	14	V2	14	V4	14	V9	14	V5	14	V3	14
Math	1298	49	.87	.08	.60	36	V4	12	V9	23	V7	33	V6	35	V2	36	V3	36	V5	36	V8	36
French	512	23	.87	.04	.65	42	V5	18	V4	28	V8	34	V3	37	V7	40	V6	41	V9	42	V2	42
Science	1144	45	.75	.14	.79	63	V6	49	V4	52	V7	57	V3	59	V8	61	V5	62	V2	63	V9	63

V1	Dependant variable Teachers estimated % Pupils achieved %	V4	Teachers' commitment to subject Major-1 ; Minor-2	V7	No. of pupils in class
V2	Sex of teacher 1 male, 2 female	V5	Marked School Certificate Yes-1 ; No-2	V8	Mean of teachers' estimated mark
V3	Teacher's service < 5 yr 1 5-10 yr 2 > 10 yr 3	V6	No subjects for Schl Cert. 5 or 6	V9	Mean of pupils' achieved mark

found between V6, V8 and V9 in Tables 4.16 to 4.19) it becomes evident that group ability is the single most important variable associated with the accuracy of teachers' predictions in the four subjects. For example, V6 appears as the first entry in both English and Science, V9 appears as the second entry in Mathematics, and V8 appears as the third entry in French. It is not surprising, in view of the correlations between V6, V7 and V8, that where one of these variables appears high in the order of entry and contributes a substantial proportion of variance to the multiple correlation, the other two variables in the triad enter later in the analysis and contribute very little further variance to the multiple correlation.

The other major variable, teachers' commitment to the subject, appears fifth in the order of entry for English (adding little extra variance), first for Mathematics and second in both French and Science (admittedly adding only 3 per cent to Science). However, the order of entry fails to reveal a reversal in direction between, on the one hand, English and Science (commitment being negatively related with teachers' accuracy) and, on the other hand, Mathematics and French (commitment being positively related with accuracy). It is difficult to attribute any single explanation to this reversal, although it may well be that both Mathematics and French attract proportionately fewer 'less-committed' teachers than the other subjects at the fifth form level and that the resultant correlation in these subjects is artificially boosted by the relative proportion of teachers in the committed and non-committed groups. (This is not to say that the values for English and Science represent a more accurate picture - see next chapter for further discussion.)

Of the remaining variables in the analysis only V5 (marked School Certificate) and V7 (size of class) play a role of any importance in the contribution of variance to the accuracy of teachers'

predictions. V5 is the first variable entered for French but the results must be treated with caution in view of the limited number of teachers involved and the fact that the same teachers typically appear in the analysis each year (that is, the data contains repeated samplings of the same teachers albeit with different groups). Further, it should be noted that for the other subjects V5 is entered towards the end of the analysis making little, if any contribution to the multiple-correlation. On the other hand, V7 is entered second in English (admittedly adding only 1 per cent), third in Mathematics (adding 10 per cent) fifth in French (adding 3 per cent) and third again in Science (adding 5 per cent). This suggests that this variable is perhaps of moderate importance in considering the question of the accuracy of teachers' predictions. It should be added that this variable may be in part related to the ability level of pupils (at least in English, Mathematics and Science - see Tables 4.16 to 4.19) and ties in with the school practice of providing smaller classes for lower ability groups.

The remaining variables, V2 (sex of teacher) and V3 (length of service) generally contribute little to the analysis in that they are entered either in the middle or towards the end of the step-wise analysis for each of the subjects.

4.5 Summary analysis and comparison of data for 1977 and 1978

The final set of results are those contained in Table 4.21 which show summary details of the accuracy of teachers' predictions across all subjects taught at the fifth form level in the years 1977 and 1978. This particular analysis was carried out in order to get some idea of the effects of feedback of previous predictions (for the year 1977) upon future predictions (1978). In other words, having informed teachers of the results from a previous year's analysis it was felt

Table 4.21. Summary of the analysis of the estimated and achieved marks of pupils in 14 subjects for 1977 and 1978.

Subject	1977					1978				
	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r	n	$\bar{x}_e$	$\bar{x}_a$	$\bar{x}_e - \bar{x}_a$	r
English	282	58	56	2	.81	328	55	57	-2	.84
Maths	257	55	53	2	.87	284	57	60	-3	.92
Science	245	53	54	-1	.81	269	54	57	-3	.87
French	57	58	62	-4	.91	61	57	58	-1	.89
Biology	18	34	38	-4	.90	22	42	40	2	.78
Ec. Stud.	56	56	56	0	.73	48	60	62	-2	.73
German	40	57	59	-2	.85	61	56	62	-6	.86
Geog.	191	50	54	-4	.85	254	53	54	-1	.88
History	117	55	59	-4	.78	140	61	57	4	.83
Latin	22	56	56	0	.89	12	65	59	6	.88
Music	-	-	-	-	-	11	62	56	6	.86
Tech. Draw	56	46	45	1	.78	71	44	44	0	.69
Typing	34	50	42	8	.68	42	55	50	5	.70
W. Work	20	46	46	0	.75	22	45	40	5	.70

that it would be interesting to see if improvements in predictions would result in the following year. It is recognized that this does not represent a true experimental design, since all teachers received the feedback from the previous year (the study did not contain an experimental and a control group of teachers - with and without feedback). For administrative reasons (i.e. for the school) it was felt more important that everyone should have feedback, that is, that everyone should be aware of the full importance of the study for the administration and provision of school staffing in future years. Nonetheless, it was hoped that the study might yield, if nothing else, hypotheses for a more detailed experimental study in a subsequent year.

In looking at the results it is apparent in fact that the feedback to teachers is not associated with any marked or systematic change across subjects for 1977 to 1978. The correlations show a slight increase in 1978 for the mainstream subjects; however, the values in a few of the non-academic subjects show a substantial decrease, for example, Biology (taken by a small group of less-able pupils over-all), Technical Drawing and Woodwork show a fall in correlation ranging from 0.12 to 0.05 (it is recognised that these values should not be seen as being linear). With respect to the difference between the estimated and achieved means of the respective subject groups, it is difficult to discern a systematic effect other than to point out that for 1978 the mainstream subject estimates were typically a few marks below the achieved marks of pupils whereas the estimates for the non-academic subjects, Technical Drawing aside, were somewhat higher than the achieved marks of pupils.

Chapter Five

DISCUSSION, INTERPRETATION AND INTEGRATION

5.1 Introduction

The major objective of this chapter is to provide interpretations, where possible, of the results obtained in this study and bring together these results with those from previous research (Section 5.2); in addition, suggestions for further research are provided towards the end of this section. The final section (5.3) provides a summary of findings from this research.

5.2 Discussion and interpretation of results

At the outset it should be stressed that the viewpoints (interpretations and comments) presented in this section are intended as possible explanations only, not formal conclusions, of the results of this study. Also, without placing undue importance upon the background experiences of the writer, it should be pointed out that the interpretations and comments provided draw upon, in part, the writer's experience of 30 years of teaching and, in some cases, an understanding of the particular nature of the functioning of the school under study.

The correlations between the estimated and achieved marks of pupils

The first point to emerge from the present study was the finding of a between-subject difference in both the general level and range of correlations for the four mainstream subjects under study. Three possible explanations are suggested for this result: the more objective nature of the content and assessment procedures in Mathematics and French than in, in particular, English; the more predictable nature of the content of the School Certificate examination in Mathematics; and the greater long-term stability of the staff of the school, and possibly their greater commitment to the subject, in Mathematics, French and Science than in English.

The first of these can be supported, at least indirectly, from research on the reliability of examinations at the 16-plus age level in Britain. Extensive studies by Willmott and Hall (1975) and Willmott and Nuttall (1975) tend to show higher estimates of reliability for subjects with a more objective criterion of measurement than subjects with a less objective criterion. The Willmott and Hall research in fact provided median values ranging from 0.70 (English language) to 0.91 (French); the next highest estimate was that for Mathematics (0.90). It does not seem unreasonable, therefore, to suggest that a more objective course content and assessment procedures is one which would enable teachers to assess more accurately, and with less variation between teachers, the ability levels of their pupils.

The second interpretation, in particular, draws upon the writer's experience both as a Mathematics and Science teacher at the fifth form level. In comparing these two subjects the writer has observed over the years a greater variability in the format and range of content of questions in Science than in Mathematics. Given support for this observation (the writer acknowledges that some may argue with this viewpoint) it is felt again that a more stable external examination pattern is likely to increase the reliability of teachers' predictions. In respect of English and French, the writer has noted a substantially greater number of adverse comments from his colleagues concerning the English examination over years (comments related to difficulty, content not covered in course and lack of predictability in the nature of the specific questions asked); in French not only are such comments less conspicuous in the experience of the writer, but also it is felt that certain components of the examination provide a more stable content from year to year (e.g. dictation given by a known teacher in familiar surroundings).

The third explanation is possibly the most tentative of the three offered. Evidence from school records shows a greater stability of staff for Mathematics and French than for English over the period of the study. It should be added that in French there is a greater likelihood that pupils would have had the same teacher for more than one year than in the other mainstream subjects. Similarly, reference to the data of this study reveals a greater proportion of teachers with first commitment to their subject in Mathematics and French than in English; however, as an explanation this is open to question in view of the step-wise regression results and an additional finding that Science teachers were at least as committed (in that they taught Science as their first subject) as French and Mathematics teachers. However, one other point should be made in relation to Science; during the course of the study the Science department experienced a change of head of department with (possibly coincidentally) an associated drop in correlation followed by an increase in subsequent years. Whilst it is difficult to suggest the precise mechanism by which a change of head could influence the results for the subject overall (there were several teachers of science in each year) it is of particular interest that the subjects with a substantial fall in correlation from 1977 to 1978 (Table 4.21) either experienced a change in head of department (Technical Drawing and Woodwork) or a change in teacher (Biology - a more experienced teacher retired and was succeeded by a much younger and less experienced teacher. Note that there is usually only one, sometimes two, classes in Biology).

The ability of teachers to pinpoint the mean level of their classes: overestimation and underestimation of marks

The second major finding relates to the differences between subjects in the values obtained for the estimated and achieved means of classes. As noted earlier, a greater accuracy was observed in

English than in, in particular, Mathematics. It must be acknowledged that it is difficult finding (or suggesting) a possible explanation for the paradox between, on the one hand, the greater ability of Mathematics teachers to order their pupils and, on the other hand, the greater ability of English teachers to estimate the general level of ability of their pupils. A tentative suggestion is that subjects with a less objective scoring scale, such as that found in English (at least relative to Mathematics and French) might well yield a more restricted mark range (smaller standard deviation) - this, at least, is the experience in British Secondary level examinations (Willmott and Hall, 1975) and thereby reduce in extent the likely difference between the estimated and obtained marks of students. It is acknowledged that such an effect is more likely to occur at the extremes of the distribution rather than at the centre.

The additional finding from the means analysis, that English teachers tended more often to overestimate their pupils' performances whereas French and Mathematics teachers tended to underestimate their pupils' performances, ties in largely with the more detailed analysis provided by the expectancy tables (and supported by the cumulative frequency graphs contained in Figures 4.1 to 4.4). In fact, the various pieces of evidence obtained yield information relating to the over- or underestimation of pupil performance at the mean, at the pass/fail boundary and across the full range of marks. The results at the mean and the pass/fail boundary are clearly, and not surprisingly, very similar. A possible explanation for the finding that the marks in Mathematics and French were generally underestimated, in contrast in particular to English, is that the content of these subjects lends itself more to rapid last minute improvements through coaching and increased pupil effort. It is difficult for teachers to estimate the full effects of such pupil behaviour. Whilst one might

expect the same pattern for Science it must be pointed out that the Science examination frequently incorporates a few questions of a type which do not lend themselves to last minute cramming; for example, pupils might be asked to explain why salt contained in a salt-pig does not become damp. This question requires a higher level of thinking (for example, analysis - evaluation in Bloom's Taxonomy (Bloom et al., 1956)) and from the viewpoint of the question content is well outside the everyday classroom experiences of pupils.

The variation in the pass/fail pattern for Mathematics across the four years (the number of passes were overestimated in 1974 and 1977 and underestimated in 1975 and 1976 - see Table 4.10) might be partly explained by a slight and unexpected shift in question content between the 1976 and 1977 examinations. In the view of the writer, the 1977 examination incorporated in places more penetrative questioning probably involving higher order thinking. Unfortunately, the data for the present study does not permit an analysis for the years 1973 and 1974 (data prior to 1974 was not kept with the same care to detail) in order to see if prior to 1974 more under- than overestimates were obtained.

The finding that teachers tended to overestimate the marks of pupils at the lower end of the scale and underestimate the marks of pupils at the top end of the scale (French excepted) is not surprising. One can envisage that teachers, by and large, might be reluctant to award marks at the extremes; the feeling possibly exists (consciously or unconsciously) that due to the restraints of the limits of the scale a highly able or low ability pupil is more likely to 'move' to the centre rather than to the limits of the scale in the actual examination setting. It should be added that the tendency to give conservative ratings to extreme cases is, as seen in Chapter Two, well documented

in the literature; Sax (1980), for example, refers to this phenomenon as the 'error of central tendency'.

The pass/fail boundary

The explanation offered earlier for the paradox between the greater ability of Mathematics teachers to order their pupils and the greater ability of English teachers to pinpoint the mean score of their classes - that less of the mark scale is used in English - might well tie in with the relatively minor (rather than substantial) differences found between the four subjects in the number of students crossing the pass/fail grade boundary. (Approximate figures of 15, 16, 17 and 18 per cent were obtained respectively for Mathematics, French, Science and English). One might tentatively hypothesise that this result reflects a likely similarity in the Standard Error of Measurement (SEM) for the four subjects. The SEM is represented by the formula,

$$SEM = \sigma \cdot \sqrt{1 - R}$$

where σ is the standard deviation of the scores in a subject and R is the reliability estimate.

The SEM thus comprises two components, σ , which is likely to be greatest in Mathematics, and R which is likely to be greatest in both French and Mathematics (i.e. greatest when considering national figures for School Certificate - see following comments). It should be pointed out that the effect of a high reliability estimate is to lower the value of the term $\sqrt{1 - R}$; thus in comparing, for example, Mathematics and English, a larger σ value in Mathematics is likely to be compensated for by a smaller value for the term $\sqrt{1 - R}$, whilst a larger value for this latter term in English is likely to be compensated for by a smaller σ value. That this might be the case both nationally and in the data of the present study is suggested by two pieces of information; first, research in Britain on O-level and CSE examinations

(at approximately the same age level as School Certificate) indicates that the SEM is relatively stable across subjects at about 6-8 marks on a scale out of 100 (Willmott and Hall (1975) and Willmott and Nuttall (1975)). Unfortunately, information on the reliability and SEM of School Certificate examinations in New Zealand is sadly lacking - the SCEB does not publish information of this sort nor is it readily accessible. As a hypothesis the writer is tempted to suggest that the reliabilities of New Zealand School Certificate examinations are likely to be marginally higher than those reported in Britain on the grounds that the ability range of pupils entered for School Certificate encompasses the ability range of both O-level (the top 20-25 per cent in Britain at the 16-plus age level) and CSE (the next 40 per cent). In general, the reliability of an examination or test increases with an increase in the ability range of the pupils taking the test or examination. The second piece of information relates to the data obtained in the present study; the correlations reported earlier between the estimated and achieved marks of pupils in the subjects under study approximate remarkably well the corresponding reliabilities of examinations at the 16-plus age level in Britain (the present values are, at worst, marginally lower). If these values are combined with the standard deviations of subjects in the present study (around 16 for French and English, 17 for Science and 23 for Mathematics - looking at the achieved marks of pupils for the four year period overall), SEMs of approximately 5.1 (French), 7.3 (Mathematics) and 8.0 (Science and English) are obtained; French excepted, these are relatively close. One further point should be made: the correlations between the estimated and achieved marks of pupils in the present study are marginally higher than those reported by either Hulbert or Gilmore, (see Chapter two) particularly in the case of French. For this reason the estimated SEM of 5.1 for French should be treated with a good deal of

caution - as mentioned, the data for French in this study is based on repeated samplings of a few relatively experienced French teachers. A more detailed comparison of the present findings with those from previous studies is given shortly (see 'The quality of teachers' estimates').

The regression analysis

The results of the regression analysis suggested that the major factor associated with accurate predictions was the ability level of the class under study. With the exception of English, the more able the group overall the higher the correlation between the estimated and achieved marks of pupils. For the purposes of interpretation, this finding is somewhat compounded by the moderate link between size of class and the triad of ability measures in the study (V6, V8 and V9). As stated earlier, it is a common school practice to create smaller classes for groups of lower ability pupils; there is also a tendency for larger classes to be given to the more senior teachers. In this context 'seniority' should not be confused with either length of service or qualifications; it may simply be that, in the head of department's judgement, the task of best preparing a large class for success in School Certificate should be given to a teacher whom the Head has particular confidence in. Confidence, in this context, is usually associated with at least a few years of common working experience between the Head and the teacher concerned and does not imply that teachers taking the smaller classes are either incompetent or that lower ability groups are being 'fobbed off' with second rate teachers. However, given that a practice exists along these lines (albeit one which is evidenced here only in the observations of the writer) the finding of a link between pupil ability and accuracy of prediction might well be partially attributable to a small variation in the

underlying quality of the teaching. Clearly, this is a difficult hypothesis to test and one beyond the scope of the present data - indeed, a number of ethical difficulties would need to be overcome before a study could be implemented to test this interpretation.

On hindsight, it could be argued that the inclusion of all of V6, V8 and V9 was unnecessary; however, the high correlations between these variables, in particular V8 and V9, suggest that the different class groups in the study were placed in some semblance of order in terms of their later achievements in School Certificate. The correlations between V8 and V9 given in Tables 4.16 to 4.19 are of a similar order to those found by Hulbert for the same subjects (see the values under the 'between' headings in Table 2.5).

For the remaining variables in the regression study, the only safe conclusion that can be reached is that any findings should be treated with a good deal of caution. Clearly, sex of teacher (V2) and length of service (V3) contributed little to the analysis; commitment to the subject (V4) produced a mixed pattern of results suggesting, at best, subject-specific links. Marked School Certificate (V5) similarly suggests, at best, a subject specific link - being of importance only in the analysis for French. Lastly, size of class (V7) whilst appearing to be of moderate importance for predicting School Certificate achievement, appears from the correlations given in Tables 4.16 to 4.19 to be compounded in part with the triad of ability measures (V6, V8 and V9). Possibly, the values obtained in Table 4.20 reflect the previously hypothesized 'quality of teacher' effect - again this is a highly speculative interpretation.

The only remaining comment to make from the regression analysis is that, in the process of the study, one or two variables emerged (untested) which might well have been of influence in the predictions of teachers. The first of these was the observation that a change in

head of department or a change in teacher might well have been associated with a lower than expected correlation between the estimated and achieved marks of pupils. Clearly, any future study should record any such events - the absence of a systematic record in this study is regretted. The other variable, associated with the variation in pattern of over- to underestimation in marks in Mathematics, is the change in examination content and format which typically occurs every few years in a subject. Again, any future study should record details of such changes or, of equal importance, changes or modifications in the underlying syllabus in each subject. Fluctuations, such as those noted in Mathematics, may simply reflect an adjustment period in the estimations made by teachers.

The quality of teachers' estimates

The finding from previous research that teachers can rank but not scale their pupils, bearing in mind the qualifications made in Chapter Two, is to a large extent supported by the present research. In drawing this conclusion one or two points need to be made; first, the correlational evidence from this research suggests that, in terms of ordering candidates, most teachers appear to be achieving results which approximate (they slightly underestimate) the reliability of the examinations themselves. If this is in fact the case (the absence of published figures for the reliability of School Certificate examinations should be acknowledged), the arguments made earlier concerning the different bases upon which teachers and examiners arrive at their assessments are, to say the least, open to question (see discussion of Petch's research in Chapter Two); it would seem here that, whatever the differences are, the order of pupils is as close as could be expected given the limitations imposed by the respective reliabilities of the external examinations and teachers' assessments. It may be

argued, of course, that the high median correlations obtained here merely indicate either that the examinations are sampling broadly and comprehensively (doing their job well) or that teachers have based their predictions on a narrow range of behaviours; whichever interpretation is taken the conclusion must be reached that examiners and (most) teachers are well in tune with each other. In this respect it is worth reviewing the present results in the light of those from previous research at the School Certificate level, namely, the studies by Hulbert (1978) and Gilmore (personal communication). With one or two exceptions (e.g. Bookkeeping/Economic Studies and Biology) the correlations obtained here are comparable or marginally higher than those obtained by Hulbert either from final school examination results or teachers' ratings. This may be seen by comparing the values given in Table 2.5 (under 'Within') with those in Table 4.21. Similarly, correlations for English, Mathematics and Science reported here are comparable (possibly marginally higher) than those reported by Gilmore (Table 2.6). There is no reason to assume, therefore, that the results from the present study are too different from what might be expected from teachers in general, although clearly more evidence nationally is still needed.

With regard to scaling, the correlations reported in the regression analysis between the estimated and achieved means of different class groups (see Tables 4.16 to 4.19) provide some evidence that teachers are relatively accurate in identifying the ability level overall of their class groups. It should be remembered, however, that although the values here correspond to those given by Hulbert for English, Mathematics and Science, Hulbert found considerably less satisfaction in the values for Geography, Bookkeeping, Biology and History, that is, for subjects not studied in depth here. Scaling, of course, involves more than pinpointing the mean level of a class;

a teacher may estimate well the overall ability level of a class group but underestimate the upper ranges and overestimate the lower ranges (as found here for English, Mathematics and Science). The evidence from the expectancy tables in fact points to the need for scaling teachers' estimates virtually at all points of the mark range, although in some cases only small adjustments would need to be made. However, given adequate criteria for scaling, the conclusion can be reasonably drawn that most teachers are able to provide as an effective (reliable and valid) assessment of pupils as the external examination itself. This comment poses one or two important questions; finding a procedure for scaling which satisfies all concerned is practically impossible, at least, if the research and comments from the present move to internal assessment are taken account of.⁽¹⁾ Secondly, in the context of validity, the argument of what is more important, the objectivity of an external examination or the more intimate knowledge a teacher has of his/her pupils' strengths and weaknesses, is one which is not readily solved by the provision of a suitable scaling scheme. The logical extension here, perhaps, is to argue for a system of assessing pupils which incorporates both an internal and an external component. This is a viewpoint expressed in Elley & Livingstone (1972) and succinctly summarized in the following comments 'borrowed' from Petch by the Schools Council (1966):

"In assessing candidates for examination purposes it seems likely that two independent assessments, one from the school excluding any forecast of the examination result, and one from the examiners, arrived at without knowledge of the school estimate, would give a better,

(1) On the topic of internal assessment, the SCEB has produced a number papers and newsletters since 1972 which make interesting reading. The reader is also referred to Schools Publication Branch, Education No. 8, 1977, for a debate of some of the issues.

(because more comprehensive), view than either assessment gives independently." (Para. 134).

Projection of sixth form class sizes

As indicated in the introduction (Section 1.1) the impetus for this research stems from the need for State Secondary Schools to know the likely enrolment figures well in advance of each new year. For the school in the present study, the estimated marks of pupils in School Certificate form an important component in the prediction of fifth and sixth form enrolments. The critical information contained in the results of this research is that presented in the individual subject expectancy tables (Tables 4.6a to 4.9d) and the expectancy tables relating the estimated number of subject passes to the achieved number of subject passes across all subjects (Tables 4.11 to 4.15). For projection purposes, it can be argued that the numbers of students crossing the pass/fail boundary (or the sixth-form entry criterion in Tables 4.11 to 4.15) is not as important as the need to obtain a balance between the percentage predicted to pass but who failed (P_f) and the percentage predicted to fail but who passed (F_p); clearly, a substantial imbalance in projections will have adverse effects on staff-student ratios or may require the appointment of a temporary teacher to counter the consequences of the imbalance. It is, of course, preferable that the size of the P_f and F_p groups be as small as possible thus diminishing the likelihood of a substantial imbalance. With these comments in mind, it is time to consider the actual values obtained.

In looking at the sixth form entry criteria (Tables 4.11 to 4.15) it is evident that more accurate predictions were made at the extremes of the ability range with a tendency for more over- than underestimates of subject passes to be made (summarized in Table 4.15). From the figures given in Table 4.15 it can be seen that the imbalance

between P_f and F_p ranges from 2 per cent (1975 and 1976) to 4 per cent (1977) of the fifth form roll. In terms of student numbers this represents an imbalance of 6 to 12 students, that is, less than half a class; a discrepancy of this order is easily coped with given that it does not mask serious imbalances at the individual subject level. The relevant information at this level has been extracted from Tables 4.6a to 4.9d and is summarized below:

Table 5.1 Discrepancy between the numbers of students expected to pass but who failed and the numbers of students expected to fail but who passed

	Minimum	Maximum
English	- 6 (1974)	-27 (1976)
Mathematics	-13 (1977)	+29 (1976)
French	+ 2 (1974)	+11 (1975)
Science	- 1 (1975)	-16 (1977)

- indicates that the number of passes were overestimated

+ indicates that the number of passes were underestimated

On the basis of the numbers provided, it would appear that the most serious discrepancies are those for Mathematics and English; in fact, given the number of classes in the sixth form for these subjects, the observed discrepancies are easily coped with by increasing (or decreasing) the class sizes by 3-5 pupils per class (alternatively, if the discrepancy is any greater, a temporary teacher could be appointed to cope with an increase in enrolments). Of greater concern is the discrepancy of 11 in the case of French, since it is a subject for which only one sixth form class is generally provided. A variation of 11 pupils clearly represents a substantial increase in the work load of the teacher concerned - any greater increase in the roll would

necessitate the appointment of a part-time or temporary teacher. It should be added that these comments apply not only to French but to several other sixth form subjects for which one class is generally provided. The remaining subject in Table 5.1, Science, is more difficult to assess since this subject becomes divided into separate fields of study in the sixth form (Chemistry, Physics and Biology). Fortunately, additional information may be applied from the subject preference indications given by pupils intending to proceed to the sixth form and in practice few problems arise in the staffing of sixth form Science subject classes (in an extreme year, the numbers for one of these subjects might be out by 6-8 pupils).

In evaluating the evidence from the four year period overall, the general conclusion must be reached that the estimates of teachers yield useful information to the administration in determining the allocation of staff to sixth form classes. It must be stressed that the preceding discussion focussed on the extreme cases in the data - anomalies of this order are in fact coped with reasonably well by the school administrator. In practice, the most serious problems arise in a single class subject when substantially more pupils pass School Certificate in that subject than was predicted.

Further research

Quite clearly, much of what has been said in this chapter represents opinion and speculation although not without reason and support. Some of the interpretations offered in this chapter lend themselves to further investigation; for example, a more comprehensive study of teachers' predictions could be made across several schools enabling the effects of subject content to be studied - do some subjects lend themselves more easily to accurate predictions? Do different assessment procedures influence the accuracy of predictions?

What are the effects of staff changes and transfer of pupils on predictions? What effects do course content (syllabus) and examination changes have on the predictions of teachers? The need to look beyond the present data is essential if generalizations are to be made nationally. In addition, it is felt that further research is needed to investigate the possibility of adopting an assessment procedure which incorporates both teachers' and external examination components. It is acknowledged that the SCEB has experimented with various schemes in different subjects in the current move towards internal assessment; unfortunately most schemes are a variation on a theme in which teachers rank their pupils within the framework of a predetermined allocation of grades (or marks) provided by the Department of Education. Schemes along those lines run the risk of failing to reward the efforts of an enthusiastic teacher working with an enthusiastic class. It is not intended to discuss in further detail here the topic of internal assessment - it is sufficient to say that the results of the present study suggest, but need confirming by a more comprehensive study, that a pooled assessment procedure will yield a more reliable and valid assessment of pupils than is currently achieved by the external examination alone.

With regard to the regression study, it is difficult to offer any substantial changes to the design which would yield more fruitful results; certainly there is a need to untangle the effects of repeated measurements of the same teachers (as in French) and to control for the effects of a change in syllabus content or a change in staff; however, it is difficult to envisage any additional variables which would be added to the regression procedure which would markedly improve the resulting multiple regression coefficients.

Finally, it is evident from the earlier discussion that the predictions of teachers yield useful information to administrators in

making projections for the following year's sixth form intake; the present study did not include a comparison of the effectiveness of teachers' estimates with other methods of projecting rolls (e.g. from historical trends over the previous 3-5 years). It may well be that other methods are available which yield essentially the same quality of information and which are less time consuming on staff and the administration; possibly two or more methods may be pooled to yield an even more accurate picture. It is felt that there is scope for research in this area: indeed, it is possible that the present study may provide ideas to researchers or administrators in other institutions where the need to make projections of future enrolments is an important requirement.

5.3 Summary of Findings

The purpose of the final section of this research is to present in one or two pages the major findings with a minimum of interpretations. The findings are listed numerically from 1 to 8 and in the main reflect the order of presentation of results, rather than an order based on importance or significance.

1. Findings from the correlation of the estimated and achieved marks of pupils indicated greater agreement for Mathematics and French (median class values of around 0.90) than English and Science (median class values of around 0.75). In addition, a substantially greater range of values was obtained for English and Science than Mathematics and French.
2. Comparisons of the estimated and achieved class means across subjects indicated greatest agreement in English and least agreement in Mathematics.
3. Results from the expectancy table analysis indicated a tendency for teachers in all subjects to underestimate the marks of pupils at

the upper end of the ability range and, French excepted, overestimate the marks of pupils at the lower end. On balance, the marks of pupils in English and Science were overestimated more often than underestimated; in Mathematics and French the opposite pattern was obtained, although the results for Mathematics showed a substantial variation across the four years of the study.

4. A study of the movement of pupils across the pass/fail boundary (in comparing the estimated and achieved marks of pupils) indicated only a small variation across the four subjects analysed in depth; figures of approximately 15, 16, 17 and 18 were obtained respectively for Mathematics, French, English and Science for the percentage of students crossing the boundary over the four years of analysis.

5. A similar analysis based on the sixth form entry criterion indicated on balance that slightly fewer pupils reached the criterion of three passes in School Certificate than expected; over the four year period approximately 13 per cent of pupils crossed the entry criterion.

6. A comparison of the level of agreement between the estimated and achieved marks of pupils for 1977 and 1978 showed no evidence that feedback of results systematically influenced the predictions of teachers; it should be acknowledged that the study in question failed to satisfy the requirements of a true experimental design.

7. Correlations between the variables included in the regression analysis of teachers' predictions indicated a factor of group/class ability comprising three measures, namely, the estimated mean of the class, the achieved mean of the class and the number of subjects taken for School Certificate; a fourth variable, size of class, was also related to this grouping of variables in Mathematics, English and Science but not French. No other systematic links between variables were observed, although a number of subject specific correlations

were noted.

8. The regression analysis yielded multiple correlations of 0.79, 0.65, 0.60 and 0.37 for Science, French, Mathematics and English respectively. Overall, the only factor to reveal a substantial and systematic link with the quality of teachers' predictions was the ability level of the class under analysis - teachers were generally more able to predict the results of their pupils for higher ability than lower ability classes (English excepted). The remaining variables showed, at best, minor links with the accuracy of teachers' predictions; these links generally failed to agree in direction (sign) across subjects. There was no evidence to link sex of the teacher or years of service to the quality of predictions.

BIBLIOGRAPHY

- BLOOM, B.S. et. al. Taxonomy of Educational Objectives. Handbook 1: Cognitive Domain. London: Longman. 1956.
- BLOOMFIELD, B. et. al. Mode Comparability in the CSE. Schools Council Examinations Bulletin, No. 36. Evans/Methuen Educational. 1977.
- BOSOMWORTH, G. The Use of Teachers' Assessments in Allocating Children to Secondary Schools. Durham Research Review 4, 1953. P.50-53. Cited in Yates and Pidgeon, op. cit. 1957.
- CHEDZOY, O.B. 'Predicted or Actual A-Levels as Indicators of Mathematical Potential?' Assessment in Higher Education, Vol. 2 1976. P.75-80.
- ELLEY, W.B. and LIVINGSTONE, I.D. External Examinations and Internal Assessments. Wellington, NZCER. 1972
- ELLEY, W.B. (Chairman). Assessment in Schools. Wellington. Department of Education. 1974.
- ELLEY, W.B. 'Overseas Experience with Internal Assessment.' NZPPTA Journal, April, 1976. P.28-32.
- FAIRBURN, K. et al. 'Internal Assessment in Queensland.' New Zealand Journal of Educational Studies. Vol. 11, No. 2, 1976. P.143-151.
- FERGUSON, C. and GARRETT, S. 'New Dimensions in A-Level History.' Educational Studies. Vol. 3, No. 1, 1977. P.11-21.
- HULBERT, M.E. Moderating Internal Assessments for School Certificate. M. Phil. Thesis, University of Waikato. 1978.
- JOINT MATRICULATION BOARD. An Experimental Scheme of School Assessment in Ordinary Level English Language: Third Report. Joint Matriculation Board, OP 31. 1970.
- LEWIS, O. 'Marks can't tell all we want to know.' Auckland Star, 17 February, 1977.

- McCLELLAND, W. Selection for Secondary Education. University of London Press. 1949.
- MARSH, R.W. Pre-entry Characteristics and University Performance. Victoria University of Wellington. 1970.
- MURPHY, R.J.L. 'Teachers Assessments and G.C.E. Results Compared'. Educational Research, Vol. 22, No. 1, 1979. P.54-59.
- NEW ZEALAND DEPARTMENT OF EDUCATION. Report on the 1969 School Certificate Examination. Bulletin No. 51. Wellington. Department of Education. 1971.
- NIE, N.H. et. al. Statistical Package for the Social Sciences. New York: McGraw-Hill. 1970.
- PARKYN, G.W. Success and Failure at the University. Vol. 1. Wellington, NZCER. 1959.
- PETCH, J.A. School Estimates and Examination Results Compared. Joint Matriculation Board, OP 21. 1964.
- QUEENSLAND DEPARTMENT OF EDUCATION. Public Examinations for Queensland Secondary School Students. Report of Committee chaired by W.C. Radford. Brisbane. Department of Education. 1970.
- RICHARDSON, S.C. 'Some Evidence Relating to the Validity of Selection For Grammar Schools.' British Journal of Educational Psychology, Vol. 26. 1956.
- SAX, G. Principles of Educational and Psychological Measurement and Evaluation. Wadsworth. 1980. 2nd edition.
- SCHOOL CERTIFICATE EXAMINATION BOARD. Internal Assessment for School Certificate. Wellington. Department of Education. 1972.
- SCHOOL CERTIFICATE EXAMINATION BOARD. Newsletter No. 3 Supplement: Modular Science. Wellington. Department of Education. 1977.
- SCHOOLS COUNCIL. Multiple Marking of English Compositions. Examinations Bulletin, No. 12, London: HMSO. 1966. Part a.

SCHOOLS COUNCIL. The 1965 CSE Monitoring Experiment. Working Paper

No. 6. Part 11, London: HMSO. 1966. Part b.

SCOTTISH EDUCATION DEPARTMENT. Secondary Education: A Report of the
Advisory Council On Education in Scotland. Edinburgh, HMSO. 1947.

SOWELL, D.J.W. 'CSE - Grades and Teachers' Forecasts.' Educational
Research, Vol. 13, No. 1, 1970. P.28-35.

SUTCLIFFE, A. and CANHAM, J.W. Selection for Secondary Education.
John Murray. 1944.

TATE, O. 'Moderation in All Things - Even Internal Assessment.'
Education. Vol. 26, No. 8. 1977. P.5-8.

WATSON, C and QUAZI, S. Ontario University and College Enrolment
Projections to 1981/82. Toronto, OISE. 1969.

WILLMOTT, A.S. and HALL, C.G.W. O-Level Examined: The Effect of
Question Choice. Macmillan for the Schools Council. 1975.

WILLMOTT, A.S. and NUTTALL, D.L. The Reliability of Examinations at 16+.
Macmillan for the Schools Council. 1975.

YATES, A. and PIDGEON, D.A. Admission to Grammar Schools. NFER
Publication No. 10. Newnes Educational Publishing Co. 1957.