THE UNIVERSITY OF HULL

An empirical investigation into Teachers' Assessments within the context of Key Stage 3 Mathematics: the development of a cognitive-model of the judgement process.

being a Thesis submitted for the Degree of

DOCTOR OF PHILOSOPHY

in the University of Hull

by

Leslie Atkinson, MA.,(Hull) BSc.(Hons.),(Durham)

October 1995.

"...despite the great deal of research already completed, it is obvious that we know very little about many aspects of information use in judgement"

(Slovic and Lichtenstein, 1971)

ACKNOWLEDGEMENTS

This investigation would not have concluded without the help and support of Dr. J.L. Moore, my Education Tutor at the University of Hull. More importantly, my efforts as a researcher would not have endured without the patience and perseverance of my wife and children.

CONTENTS PAGE

LIST OF CONTENTS

CHAPTER PAGE

LIST OF FIGURES PAGE

Chapter 1.

Professional Judgement: a problem within secondary education.

1.1 Introduction.

There are a variety of interpretations of the term 'professional judgement'; some precise others less involved. However, all interpretations (should) address the general problem which is concerned with the integration of information to produce a decision (or judgement). Although evident within a range of, mainly vocational, occupations, it is perhaps within the fields of medicine, education and the judiciary that the notion professional judgement is most familiar. In-spite of this high profile and the research interest, reported within the literature, very little is still known about many aspects of information use in judgement.

It is probably within education that a specific need for research into professional judgement can be most readily identified. Over the past three decades there has been a gradual change in established educational practices with an increased emphasis on the academic assessment of students directly by teachers. Inevitably, this has brought into focus the judgements undertaken by teachers whilst performing such assessments. Initially, the educational initiatives associated with the promotion of professional judgement were of a

relatively small scale. Often these initiatives would be directed at a particular phase of education (for example, GCSE coursework). However, towards the latter part of the 1980s, the term professional judgement became firmly embedded within the working vocabulary of teachers at the levels of primary and secondary education.

1.2 The need for the research.

The introduction of the National Curriculum in September 1989 required a formality to the role of 'teacher assessment' hitherto unheard of within the educational profession. The additional responsibility allied to the proposed high profile approach to be adopted for professional judgements caused some concern to both teachers and educationalists alike (HMI, 1989). The essence of this concern centred around the inherent complexity of the assessment framework on which the National Curriculum was founded. The use of a multitude of criteria, or statements of attainment as they were to be called, describing what pupils know, understand and can do, bore the brunt of the criticism echoed across a variety of educational literature, ranging from the Times Educational Supplement (Mortimore, 27th July, 1990) to the Curriculum Journal (Murphy, 1990). The depth of feeling is (probably more readily) summed up by the comment made by Gipps (1990), where criticism of the National Curriculum assessment framework is made in the context of the

failed introduction of a similar curriculum initiative in Scotland in the mid 1980s:

> "With the evidence of the Scottish experience in mind, there is no doubt that the system as it is proposed is unworkable."
>
> (Gipps 1990, p96)

The concerns voiced over the general nature of the assessment framework were similarly felt at the individual subject level. The history of assessment expertise compiled over recent years within the secondary school mathematics fraternity, for instance, allowed for a focus of critical comment from this quarter. An influential body within secondary mathematics, the Schools Mathematics Project (SMP), was quick to provide both guidance and advice for those schools using SMP schemes. This information detailed not only resource materials to facilitate the 'teacher assessment' elements of the National Curriculum, but also gave persuasive interpretation of the methods to be employed by teachers to allow a 'realistic workability' to be achieved within the proposed framework. The SMP's view of the statements of attainment, which are themselves the essential building blocks of the whole assessment process within the National Curriculum, is made quite clear from the following:

> "It is arguable that these individual statements cannot in themselves be precise criteria by which to measure pupil attainment. Trying to 'tick off', one by one, each statement once 'sufficient' evidence has been shown that pupils have 'achieved the statement' would then be an exercise of doubtful validity, since the statements, taken by themselves, are not sufficiently well-defined."
>
> (SMP, 1990, p5)

The general concerns referring to the introduction of the National Curriculum tended to centre upon issues of complexity and workability. These were by themselves 'vague' and lacked the specifics of problem identification or qualification. However, this was undertaken and achieved within Mathematics, highlighted by the deliberations of the SMP. Hence, it is within the context of the Mathematics National Curriculum that the problems associated with 'teacher professional judgement' appear to have been most clearly defined.

1.3 Aim of the thesis: an overview.

The principal aim of this thesis is to investigate professional judgement within the vocational context of education. More specifically, the Mathematics National Curriculum at the secondary school level provides the necessary and appropriate focus for a researchable problem to be defined and explored. The documented investigation may be catalogued within three distinct aspects. Each will now be described in brief.

Firstly, an outline will be provided of the broader issues associated with educational assessment. this will indicate the extent in growth of the demands upon professional judgement within teaching over the past thirty years. Educational initiatives highlighting the importance of teacher assessment will be identified and related to the judgement process in

particular. The issue of professional development will then be considered with specific reference to the implications for the In-Service Education and Training of teachers. This initial review of the literature forms the basis of chapters 2 and 3.

Secondly, the issue of professional judgement will be considered within the specific confines of those particular educational initiatives leading up to the introduction of the National Curriculum. The review of the National Curriculum will be followed by the reporting of a small scale interview schedule focussing on those areas of concern mentioned within the earlier literature review. The interview schedule is directed towards specific issues associated with the Mathematics National Curriculum and teacher assessment. From the review of literature together with the reporting and analysis of the interview schedule it was possible to delineate a researchable problem in terms of a series of stated aims and testable hypotheses. This work is documented within chapters 4 and 5.

Thirdly, in order to investigate the outlined problem a research design was adopted, test-instrument and survey questionnaire developed, and administered. The procedural aspects of both the pilot and main study versions of the research design are described and analysed utilising a comprehensive series of statistical techniques. The documented

development, administration and analysis stages of this study form the basis of chapters 6 and 7.

The final chapter of the thesis will summarise the analysis results, discusses the findings and provides an evaluation of the study. A consideration of the implications of this research for the field of professional judgement within education (and otherwise) will conclude this chapter and the thesis.

Chapter 2.

Assessment: A review of purpose and practice.

The aims of this chapter are to outline the primary purposes and practices of assessment. It will illustrate the evolving nature of assessment and consider in detail certain aspects of the more important educational measurement techniques. School Based Assessment will be discussed. This discussion will address the concerns surrounding the reliability and validity of School Based Assessment and its relationship with Criterion Referenced Measurement.

2.1 Introduction.

The 'new teacher' to the profession has a 'high expectation' of assessment as a means of collecting a range of information and data relating to numerous aspects and characteristics of his or her intended charges. This anxiety is expressed by one such 'new teacher' in the following extract:

> "Jan recognized that if she was going to be successful in her new job, one of the things she would have to do would be to obtain a great deal of information on her pupils. Only then would she be able to make valid decisions about how to plan her teaching"
> (cited in Lindvall & Nitko, 1975, p3)

Assessment, in practice, may well dominate the teaching and learning process for the duration of a teacher's academic career. The notion of 'high expectation', within the Mathematics curriculum at least, may account, in part, for the

increasing level of discussion and debate centred around assessment in schools over the past few years (Noss et al, 1989). The influential Cockcroft Report (1982), acknowledged the importance of assessment, but indicated variety and purpose should be key features of any adopted procedures.

2.2 The purposes of assessment.

The purposes of assessment are many, although they may be categorized into those which generate information and data for internal school consumption and those which do so for school-external destinations (Ahmann & Glock, 1975). This dichotomy of purpose is but one interpretation. However, it does provide a basis and starting point for discussion regarding the role of the teacher within the assessment process. As Lindvall and Nitko (1975) have pointed out:

> "The essential purpose of teaching is to produce changes in pupils.....the degree of teacher success can be determined only through regular assessments of what pupils have learned."
> (Lindvall & Nitko, 1975, p4,5)

This identifies a theme on which Lindvall and Nitko focussed; that of evaluation. It is quite feasible for the collection of information and data through pupil assessments to allow for two distinct forms of evaluation. The distinction between the evaluation of pupils and that of teachers can, in some circumstances, become a matter of the perspective placed upon

the assessment process utilized. This perspective will now be looked at in more detail.

Assessment, as previously mentioned, has many purposes, although there is one which is probably more readily identifiable to most practising teachers. Thorndike and Hagen (1977) indicated the importance of the assessment process as a means of "informing day-to-day decisions of the classroom teacher" (p166). The very process of teaching and learning is seen to depend on the evaluation of assessment outcomes as an integral part of the teacher-pupil interaction. Chase (1978) also viewed assessment as an integral part of the teacher-pupil interaction, but preferred to delineate its purpose in terms of behavioural changes. It is with reference to the size and direction of these behavioural changes that informed educational decisions may be made.

In contrast to the above, Desforges (1989) characterizes the purpose of assessment with a broader view:

> "It is generally held that the main purpose of assessment is to provide information to help people make decisions..... pupils, teachers, parents, employers and local and national policy-makers all make educational judgements."
> (Desforges, 1989, p3)

This broader based view, may well reflect the change of perspective placed upon assessment during the 1980s. The assessment process, in this context, provides a range of evaluators with information, not just the teacher. Each

evaluator is likely to have a differing use or need for the assessment data gathered. The teacher, for example, will be predominantly concerned with the planning of what is to be taught or learned on a daily basis. The national policy-maker, however, would be concerned with longer range financial and gross detail curricular planning.

The notion of assessment 'serving more than one master' is not in itself new, Desforges (1989) has identified the following six possible purposes of assessment: (i) diagnosis, (ii) evaluation, (iii) guidance (iv) grading, (v) selection and (vi) prediction. More recently the purposes of assessment have been expressed, by Gipps (1990), also in terms of six key features. These are briefly detailed as follows:

(a) screening - a process of identifying children with specific educational needs, these include those pupils referred to as special educational needs;

(b) diagnosis - a process of identifying particular strengths and weaknesses of individual children, sometimes involving the use of standardized tests;

(c) record keeping - the storage of test and other assessment information to allow for teaching and learning decision making to be undertaken and for use by other interested parties (parents, colleagues, other educational institutions);

(d) feedback on performance - the analysis of assessment

data to ascertain the level of progress and success of pupils, for use by the teacher and other interested parties (Head of Department, Headteacher, Education Authority, Department for Education);

(e) certification - externally (or internally) accredited qualifications provide a record of performance relating to a specific level of competence or knowledge within a given domain of study;

(f) selection - the utilization of information for the categorisation or allocation of children in terms of their suitability for different teaching groups (streams/bands), other educational institutions or employment.

Essentially, these six uses can be classified as to their professional or managerial nature. Professional, in this respect, means the extent to which the purpose aids the teacher in the process of educating the pupil. In contrast, managerial involves the use of assessment data to aid the management of the education system overall. It is evident from a consideration of each purpose that some fit more readily into one or other of the two classifications. However, 'feedback on performance' and 'record keeping' are seen to be part of both.

Although it is clear that assessment, including testing, plays a key part in the day to day practices of the classroom teacher

over the past decade this role has evolved to encompass a further function, this function being of greater importance than it was to the traditional role of evaluation. Gipps (1990) indicated a broader view of the purposes of assessment which depicted aspects and uses beyond the classroom. It is the emphasis and importance placed upon the evaluation of assessment data which has changed. Once assessment had as its main beneficiary the pupil - a professional use; more recently evaluation has centred upon teachers and schools - a managerial function.

2.3 Educational measurement: two contrasting styles.

The formalization of assessment, or 'educational measurement' as it was referred to by Glaser (1963), probably first took shape in the early part of the 1960s. This formalization produced a dichotomy within the broad feature of achievement measurement. This division of 'thought' and subsequent 'practice' yielded contrasting educational measurement techniques namely Norm-Referenced Measurement (NRM) and Criterion-Referenced Measurement (CRM).

Norm-referencing evolved from the need to interpret the raw score an examinee obtained on a test with greater clarity or meaning. Nitko (1980) suggested an examinee's performance could be better interpreted if the raw score were compared with, or

referenced to, something other than the test itself. In essence, an individual examinee's score would be compared to the performance of a larger group or 'norm' of examinees. The use of statistical techniques of analysis and data adaptation are invariably associated with such assessments, providing the necessary reference scores or performance standards.

Criterion-referencing, in an analogous way to that of norm-referencing, evolved out of the need to interpret better an examinees raw score on a test. The emphasis in this respect was not on the comparison of performance against other examinees but to that of pre-determined behaviours or in terms of specified performance levels. Glaser (1963) indicated the basic principle of criterion-referenced measurement to be the judgement of the individual against a continuum of inter-related behaviours. The process of testing allowing the position on this continuum to be determined.

These statements of purpose present well established versions of the basic concepts underlying each form of assessment. Norm-referencing as a concept is probably more easily understood through its familiarity of use within educational circles. Although, the interpretation of 'norm' varies form source to source. A precise definition has been afforded by Popham (1978) where performance on a norm-referenced test is judged against a group or cohort of students whose results form a series of

reference points, the cohort usually having taken the test at some previous time. William (1992a), however, considered 'norm' to mean the average for the group of students taking the test to which the individual's result is compared. The term group could relate to a teaching group, school population, local or national examination cohort. Mobley et al (1986) in their review of GCSE examination practices supported this view. This particular interpretation has through its perceived association with CSE/GCE and latterly GCSE grading practices become the accepted definition of norm-referencing. Criterion-referencing, however, is more problematic in its interpretation.

Gray (1978) has indicated the magnitude of the misinterpretation of the concept of criterion-referencing through a comprehensive review and analysis of its definition. The result of a content analysis of 57 references illustrated a broad division of thought regarding the basis on which criterion-referencing is founded. The majority of the definitions described, implicitly or otherwise, the notion of a domain of behaviours to which performance is referenced. The remaining definitions involved, again implicitly or otherwise, the notion of a continuum of behaviours to which an examinees performance would be referenced. The distinction between the two definitions is an important one, and will now be considered in a little more depth.

In both definitions the practice of educational measurement is essentially similar - examinees undertake a test, the results of which are referenced to some external set of values or scores (external to the test). It is the nature of the referent which distinguishes the two definitions. A 'domain' may be thought of as a set of related behaviours which are for all practical purposes unorderable along any specific dimension. Tests of this nature may be thought of as 'domain referenced' tests (e.g. Popham, 1969). In contrast, a 'continuum' may be considered to be hierarchical in nature, allowing for the particular set of related behaviours to be ordered (scaled) along a specified dimension. Tests which purport to reference to a continuum are often further sub-categorised by the nature of the behaviour relationships within the ordering itself.

It is the context of the learning process and its relationship with the referent which provides the main difference between the two main variations of continuum referenced tests. Firstly, the hierarchy may be considered as a series of learning stages whose ordinality is determined only by the degree of difficulty or complexity of the behaviours within the continuum. Similarly, the hierarchy may depict a sequence of learning which is of a prerequisite nature - each stage of learning facilitating positive transfer to the next. The difference is a subtle one but important as ordering behaviours on the basis of

difficulty or complexity does not necessarily lead to a verifiable prerequisite learning sequence - something often desirable but not always possible to accomplish. A notable example of a continuum based on degree of difficulty or complexity is illustrated within the Concepts in Secondary Mathematics and Science (CSMS) research (Hart, 1980). The principal aim of the research was to develop a hierarchy of levels of understanding. The hierarchy was ordinal in that the theorized conceptual levels were arrived at empirically using a simple 'this is harder' criteria. The actual hierarchy developed within the CSMS was found to be essentially uni-dimensional and also 'surprisingly robust' when used in practice (Brown, 1989, p126). Although confidence in the validity of such hierarchies have been called into question; Vergnaud (1990) indicated caution in dealing with conceptual relationships in any simplistic manner. There are several examples which illustrate prerequisite learning hierarchies, probably the most notable is that of Gagne (1968). Gagne envisaged a learning hierarchy as:

> "an ordered set of intellectual skills such that each entity generates a substantial amount of positive transfer to the learning of a not previously acquired higher-order capability"
>
> (Gagne, 1968, p3)

Although a prerequisite learning sequence appears to imply the existence of a unique path or route through which the learning process develops; this is not necessarily the case. Work by

Gagne et al (1962) centred upon the acquisition of particular mathematical skills indicated such learning hierarchies to be of a multi-linear nature.

The contrast between the referencing of examinee performance to a domain and a continuum of behaviours, in conjunction with the subtlety within learning hierarchies, indicate a potential source of misinterpretation for those defining CRM. To exemplify the problem of misinterpretation consider the frequently quoted definition of criterion-referencing, offered by Glaser and Nitko (1971). The nature and content of this definition are seen to cause confusion over the issue of a continuum of behaviours, a feature which was prominent in Glaser's (1963) original definition. The definition of criterion-referencing is expressed as follows:

> "A criterion referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Performance standards are generally specified by defining a class or domain of tasks that should be performed..... measurements are referenced directly to this domain for each individual measured"
>
> (Glaser & Nitko, 1971, p653)

As Gray pointed out the definition lacks an acknowledgement implicit or otherwise of any continuum; yet when placed in the context of the entire work of the authors a continuum is found to be implied.

In order to clarify the existing definitions of criterion-referencing, Gray presented the following definition for this technique of assessment:

> "Criterion-referenced tests are those designed to produce measurements directly interpretable in terms of specified performance standards where the standards form a continuum of knowledge that is dependent on the prerequisite relations among the various levels of the continuum."
>
> (Gray, 1978, p227)

Gray, additionally comments on the delineation of the concepts of a continuum and prerequisite relation within the proposed definition by the following comment:

> "It in no way excludes the idea that a continuum may be multi-linear, and it assumes that a prerequisite relation is one in which the lower-order competency is a prerequisite because it promotes mastery of the higher-order competency through positive transfer."
>
> (Gray, 1978, p227)

Gray's definition is more restrictive than most associated with this technique of assessment but it does have the advantage of combining Glaser's (1963) original notion of a continuum with the Glaser and Nitko (1971, p653) definition of criterion-referencing. In addition, the tacit acceptance of Gagne-type prerequisite relations, accommodated within a multi-linear continuum counters Vergnaud's (1990) over-simplification charge. Consequently, Gray's (1978) definition of CRM will be that utilized within this study.

2.4 Assessment: issues of reliability and validity.

Before describing and illustrating the various aspects of reliability and validity, it is important to place in context the concept of educational measurement within the field of assessment. For the purpose of this study an educational measurement may be thought of as the interaction of a particular test instrument with its designated subject(s) (Popham, 1978). Normally, the test-instrument will be in the form of a criterion or norm-referenced Test, yielding CRM or NRM. It is the nature of the test-instrument which ultimately determines the utility of the assessment processes outcome. This notion of a test instrument's effectiveness is commonly expressed in terms of associated reliability and validity measures.

The concepts of both reliability and validity in connection with test instruments in education have been defined on numerous occasions by many authors. Unfortunately, as Ahmann and Glock (1975) have pointed out, the concepts of validity and reliability have so often been used as though they were synonymous; when they are not. Possibly some of the confusion surrounding these 'twin' concepts may be due to their interdependence. As Nuttall and Willmott (1972) comment, reliability is a necessary but not sufficient condition for validity. Both concepts will now be dealt with briefly.

Chase (1978) has provided the following definition of reliability:

> "A test is reliable to the extent that it is consistent with itself, that is, it ranks the individual in essentially the same position on successive applications."
>
> (Chase, 1978, p79)

The important feature of this definition is that of self-consistency. This concept provides the basis for most reliability definitions. In contrast, Thorndike and Hagen (1977) perceived reliability more technically in terms of the accuracy and precision of a measurement procedure. Precision in this context refers to each individual measurement; and accuracy relates to the level of reproducibility of each individual result. From reviewing the literature three distinct aspects of consistency, each yielding a different type of reliability, emerge and these are used predominantly within educational measurement. Additionally, William (1992a), in a more recent, comprehensive review of some technical issues in assessment, includes the less often cited mark-remark reliability associated with teacher assessment. All four aspects of consistency and types of reliability are illustrated in Figure 2.1. Associated with each reliability measure are statistical techniques, which allow these to be expressed (quantitatively) as reliability coefficients. A statistical treatment of reliability is given in the data analyses sections of this study (chapter 7).

Figure 2.1. Aspects of consistency and types of reliability.

CONSISTENCY	RELIABILITY
If the same student was assessed on two occasions using the same test, would the marks be equal?	test-retest reliability
If two equivalent forms of a test were used would the same student get equal marks on both tests?	parallel forms reliability
If the test were split (into equal sections) would the student obtain equal marks on all parts?	split-half reliability; internal consistency reliability
If the same student was assessed by two different teachers, would their marks agree?	mark-remark reliability

The concept of validity, as with reliability, has been the subject of numerous definitions. Hoste and Bloomfield (1975) provided a definition which probably expresses the essence of this concept most concisely:

> "The validity of any assessment procedure is determined by the extent to which it measures what it sets out to measure."
>
> (Hoste and Bloomfield, 1975, p21)

It is Borg and Gall (1983) who have commented on the key feature associated with all validity measures, that of their fitness for purpose. A test instrument may be quite valid for one purpose, and yet invalid for another. A review of validity definitions identifies the existence of three distinct aspects providing an equivalent range of validity styles. Again William

(1992a) identifies a further style, rarely cited, that of backwash validity. All four aspects of purpose and types of validity are illustrated in Figure 2.2.

Figure 2.2. Aspects of purpose and types of validity.

PURPOSE	VALIDITY
To determine the extent to which the test content is a representative sample of the total content domain.	content validity
To determine the matching of students' test scores with those obtained on another test of relevance.	criterion-related validity;
To determine the extent to which test results correspond with predictions based on a psychological theory.	construct validity
To determine the extent to which any assessment procedure interferes with the process of validation.	backwash validity

The extensive depth and breadth of literature associated with the issues of reliability and validity indicate these are of considerable importance to the educational researcher. The use of these two measures with CRM and NRM tests has been well documented. It is arguable, though, whether due recognition has been afforded the differences between these two types of test and the consequence(s) this may have on their respective relationships with reliability and validity measures.

Carver (1974) suggested that although the eligibility of reliability and validity measures for use with CRM and NRM was not in dispute, the interpretation of any resultant data or information is questionable in certain circumstances. The basic kernel of this argument is that outcomes of reliability and validity measures need to be interpreted differently for NRM and CRM tests. It is clear that differences of analysis are required for both tests as each tends to focus upon a differing aspect of measurement. NRM attempts to identify and exploit the 'between-individual score differences' found when a group of examinees undertake a test; while CRM attempts to highlight the 'within-individual score gains' of examinees having undertaken a test. Carver contrasted the work of CRM with that of the experimentalist in the physical sciences, intimating that parallels needed to be drawn between the application and interpretation of any respective data analytic techniques. Although Carver may well have oversimplified the debate regarding the inappropriate use of phychometric data analysis methods, commonly associated with NRM, his cautionary approach to the interpretation of CRM information, after the application of such techniques, merits consideration. This is given in the later data analysis section of this study (chapter 7).

2.5 School Based Assessment: an assessment alternative.

It is the relationship between assessment and the curriculum which appears to be a common theme of the literature within the field of School Based Assessment (SBA). There is an expressed anxiety that external assessment may 'drive' the curriculum and not take its rightful place in partnership within a suitable curricular-framework. The nature of this anxiety and the context of the curricular-framework are both illustrated within the following extract from Mathematics 5 to 16:

> "...it is important to emphasize that assessment should develop out of the curriculum, its aims, objectives, criteria for content and approaches, and not the reverse. Neither narrow assessment techniques nor cluttered examination syllabuses at 16-plus should be allowed to distort the aims, objectives and approaches required in a mathematics curriculum which is broad, balanced, relevant and suitably differentiated."
>
> (DES, 1989, p44)

SBA provides, in part, a framework, rather than a solution, within which assessment and the curriculum can become a partnership. A key element of this framework is the importance placed upon the teacher now directly involved within the assessment process. The move from centralized assessment arrangements (for the 16+ examination) to its partial devolvement to schools (and individual teachers) has been implemented with an understandable reluctance on behalf of many governments. It has long been realized by educational policymakers that assessment and its resultant test data could

be used as an administrative mechanism for the implementation of educational policy (Madaus, 1985).

The desire for SBA generally has been advocated for quite some time. The initial excursions into SBA were during the early 1960s with the introduction of the CSE Mode III examination. The ensuing thirty years witnessed SBA gathering momentum in terms of both prestige and credibility - becoming a compulsory element of all GCSE examinations in 1987. The introduction of SBA as a compulsory element of GCSE was an acceptance of the limitations of the then current external 16+ examination system, although such a view had not been a new one. Within mathematics Cockcroft (1982) acknowledged that timed-limited written papers could not assess all aspects of mathematical ability. Caplan and McAfee (1977) echoed this sentiment, albeit some time earlier; insisting that a variety of information gathering processes should be employed during the act of assessing. Subsequently, Buckle and Riding (1988) have provided additional support for this view which indicated that a considerable body of research had shown many pupils were predominantly verbal in their preferred mode of response - a mode not accommodated within time-limited written examinations.

It is not just critical comment regarding the limitations of traditional assessment techniques which have been the centre of attention for the SBA advocates. The positive qualities of SBA

have been variously cited in the argument for its adoption. For example, Better Mathematics (HMSO, 1987) commented on firstly, the potential for 'richer assessment', which is possible through classroom-based activities, and secondly, the subsequent enhancement of teachers' sensitivity and confidence of their assessment abilities providing credibility to such professional judgements. In conclusion, it is probably the Secondary Examination Council (SEC) who have contributed the most important reason for the adoption of SBA, that of curricular validity. It was genuinely felt that many examinees were unable to demonstrate their true abilities within a subject through the use of time-limited written papers alone. As the SEC (1986) have pointed out:

> "teachers at present usually undertake school-based assessment because they perceive its curricular benefits or because they consider it a 'fairer' form of assessment."
>
> (SEC, 1986, p2)

The major problem SBA had to confront through its development was that of credibility. The compulsory, partial devolvement of assessment from the Examination Boards in 1987 was not coupled with a commensurate reduction in their degree of responsibility for ensuring standards. It was therefore necessary for the Boards to improve their support provision for SBA and to consider alternate forms of assessment which were both reliable and valid (Luijten, 1991). The need to explore assessment techniques which were capable of delivering the broad range of

educational measurements associated with SBA brought an inevitable attention upon CRM. This attention did not merely centre on the SBA component of the evolving 16+ examination but was also concerned with the time-limited written aspect. Consequently, the more traditional assessment techniques became less dominant in the latter part of the 1980s. Norm-referencing as a concept was replaced with a Criterion-related assessment and examination grading system within the new GCSE (Johnson, 1989).

2.6. School Based Assessment: an alternative in practice.

The development of School Based Assessment has gradually seen the utilization of a variety of techniques in the assessment of pupils by teachers. The need for variety has been essential to allow for assessment to be undertaken beyond the boundaries of that encompassing mere facts and skills, covering areas of conceptual structures and general strategies (DES, 1987). Pencil and paper time-limited testing has had to give way, in part, to other methods of assessment capable of dealing with a more diverse range of response formats: oral, aural, written, practical and more recently microcomputer based. Whatever the response mode employed by students during any assessment it is, as Pirie (1988) pointed out, vital that teachers' have confidence in their own abilities to assess pupils within the normal classroom environment. This professional concern with

issues of reliability and validity is of equal importance from the managerial viewpoint of assessment legitimacy.

SBA as a credible method of assessment was recognized as a viable alternative to traditional assessment as early as the mid-sixties. Research into the use of this technique, undertaken at the time, concluded teachers were capable of providing assessments that were as valid and reliable as those associated with more traditional, externally produced examinations (Schools Council, 1967a, 1967b). Several other, more recent, case studies of SBA in practice (Torrance, 1986, Bain, 1988, Hayes, 1991) indicated that this style of assessment had significant curricular advantages over more conventional forms. These advantages ranged from: possible application for the assessment of classroom-based groupwork, to potential for use with pupils having special educational needs; both extremes normally impossible to accommodate within a more conventional assessment framework. Such diverse assessment opportunities allowed within SBA are further complimented by its merits, when judged directly against more conventional forms of assessment. Croskery (1988) produced substantial evidence to support the claim that SBA could increase an examinee's level of academic achievement.

It is probably the ability of SBA to make eligible for assessment those 'everyday' aspects of classroom practice,

groupwork for instance, that is so appealing in the view of the teaching profession. The possibility of recording, for the purposes of an external examination, key evidence, which may be ephemeral, displayed within the normal classroom environment appears to give this form of assessment a significant advantage over more traditional ones.

Although SBA has many advocates, it is not without its critics. Case studies of SBA implemented within Australia as a comprehensive alternative to external examinations have not been received positively by all involved. McBryde and Lamont (1980) reported concerns regarding the comparability of internal examination standards between schools. Findlay (1987) questioned the cost of SBA and suggested little educational progress had resulted from its introduction.

Coursework and practical work are probably the most obvious manifestation of SBA within the current 16+ examination system. These aspects have themselves been criticized for a variety of reasons. Lord (1987) itemizes several concerns regarding coursework in mathematics. These concerns ranged from the need for key administrative and organizational facilities within schools to the requirement for growth in the professional assessment standards of teachers, all deemed necessary if coursework demands were to be adequately fulfilled. Similarly, Moore (1989) found, during a study of SBA within thirty schools

(in Northern Ireland), that policy and practice did not always match. Again issues of administration, organization and professional concerns relating to assessment standards featured within the findings of the study.

It is difficult to ignore the criticism levelled at SBA when much of this is substantiated by research evidence. However, the very nature of this assessment style will nonetheless be always open to criticism because of its dependence on teachers' subjective, professional judgement. Additionally, the symbiotic relationship between SBA and CRM, which has evolved over the past few years, is equally likely to promote an unease towards this aspect and style of assessment due simply to teachers' lack of familiarity with its use.

2.7 Discussion.

The aims of assessment are seen to be essentially two fold: managerial and professional. It is within the political arena that the managerial aspects reside on the whole. The need and desire to control the curriculum, through assessment is one facet of this. The professional aspect of assessment is that which most concerns teachers. The late 1980s witness a growing awareness that assessment should really conform to the notions of 'fitness for purpose' (SEC, 1986) and the broader aim of

'making what is important measurable rather than what is measurable important' (Mobley et al 1986).

The objectives by which these aims could be attained were also two fold: CRM and SBA. It was Fremer (1972) who indicated the potential flexibility of CRM as an assessment technique. In many respects CRM, in the early years, was a solution without a problem. The need to increase the curricular validity of external examinations together with the emergence of CRM, as a credible assessment technique, allowed SBA to become a reality. However, the abandonment of NRM was seen by some to be a little premature. Fitzgibbon (1972) contrasted the position of CRM and NRM styles of assessment within the context of educational measurement and concluded both have a part to play.

It is interesting to speculate which of CRM and SBA is the driving force of the partnership. Possibly their symbiotic relationship is simply one of solution and problem respectively. Irrespective of the outcome of such a hypothetical debate, their intertwined relationship is one difficult to disentangle and is essentially complex. The unfortunate fact is the developments in educational assessment since the mid-sixties have, in general, caused a degree of confusion to teachers and other educationalists (Gipps and Goldstein, 1984). The relationship and position of CRM within the external and internal examination system appears to be a

focus for this confusion. Although as Robinson (1988) commented, the reforms undertaken during this period have integrated assessment and curriculum issues within schools, they have also promoted the greater involvement of teachers in the assessment of pupils. Torrance (1988) however, did not view this expansion of teacher involvement as necessarily good or desirable:

> "The involvement of teachers in school-based assessment per se - marking work under instructions from examiners who in turn are ultimately operating under instructions from government - is clearly not the same thing as school based-examining - the design and assessment of courses within the school."
>
> (Torrance, 1988, p34.)

Although Torrence expresses a technically correct objection to the realities of SBA, it is probably the more pragmatic view of Johnson (1989) which projects the prevailing attitude, in general, of the teaching profession towards this form of assessment:

> "Whether conscript or disciple the fact is that internally assessed work (and the external moderation which accompanies it) are here to stay, and are an important and integral part of the public examination system at 16+."
>
> (Johnson, 1989, p1)

Whatever the position or view taken on this subject there are clear benefits to be gained from the increasing involvement of teachers within SBA. Not least of which this involvement must provide a situation promoting greater debate and discussion over the nature and application of particular key educational aims and objectives and their means of fulfilment.

2.7.1 Summary.

The growth of SBA over the past thirty years - to its current status as a compulsory element of all GCSE Examinations - has been surrounded by the inevitable questions of reliability and validity from the educational traditionalists. Although these questions have been answered to the satisfaction of the Examination Groups and also the Government there are still the on-going concerns relating to standards. In particular the monitoring and ensuring of standards raises several questions. The ability of teachers to perform reliable and valid assessments within SBA is the focus of much of this attention and concern. Within this context it is the 'professional judgements' undertaken by teachers whilst performing school-based-assessments which are important. The next chapter will review the literature relating to 'professional judgement' within SBA. The review will encompass assessment reliability and validity together with issues of accountability and In-Service Education and Training (INSET).

Chapter 3.

Assessment by Teachers: A review of professional judgement.

The aims of this chapter are to describe the nature and function of professional judgement within the broader context of School Based Assessments undertaken by teachers. This will involve reviewing some of the important and more pertinent studies featuring professional judgement in educational practice. Finally, professional judgement will be discussed more generally. This discussion will encompass the key issues of assessment accountability and In-Service Education and Training and their relationships with quality control and quality assurance.

3.1 Introduction.

The conventional approach to educational measurement involves the use of a test-instrument, normally in the form of a criterion or norm-referenced test. The assessment process in this context generates a CRM or NRM. Within the professional judgement process, however, the teacher is asked to assess a subject by means of a decision (often referred to as a rating within the literature). Therefore the teacher becomes the test-instrument (van der Kamp, 1976). In this context, it is the decision strategy *used* by the teacher (or rater) which determines the outcome of the judgement (or rating) process. In

practice, the numerous teacher-pupil interactions involved with such measurements may well be interpreted with reference to a series of judgement (or rating) criteria or to a normative group or cohort (Harding, 1989). The professional judgement process, therefore, may be thought of as a distinct but subsidiary element of School Based Assessment.

3.2 Professional Judgement: a practice in need of a theory.

The unique characteristic of professional judgement, which differentiates it from more conventional assessment techniques, is specifically its reliance on the teacher as the test-instrument. This reliance, by implication, introduces an appreciable degree of subjectivity to the measurement process, often involving judgements based upon a collection of evidence, some of which may be ephemeral. Although such judgements are possible on a variety of student characteristics, a preliminary review of the relevant literature tends to indicate teacher expectation of academic performance to be the central focus of many educational studies (for example, Hoge and Butcher, 1984). Consequently, this aspect will be the initial and prominent feature explored within the broad subject area of professional judgement in (educational) practice.

The practice of professional judgement is not in reality the simple one-way interaction implied by the term. The judgement

process is in effect a true form of communication (two-way) between the teacher and the student. This notion was commented upon by Cooper et al (1982) indicating that teacher expectation of a student's academic performance can effect the behaviour of the student and in turn the teacher. However, the notion or concept of expectation is itself a question of some concern:

> "It is also likely that expectation effects are dependent on how expectations are defined. Although a number of definitions of teacher expectations have been employed in research, we know little about their explanatory value."
>
> (Cooper et al, 1982, p577)

Later, in a critique of seven studies, Hoge and Butcher (1984) suggested that teacher preconceived ability expectations of students can have an effect on their predicted achievement ratings - supporting Cooper et al's (1982) theory. The notion of test results, artificial or legitimate, having an effect on teacher judgements is not new, and is variously cited as the 'halo effect'. Owen (1976) indicated the halo effect to be a key problem associated with teacher judgements. Airasian (1977) found, in a study of 47 teachers and 1566 students, that some teachers raised their achievement expectations once key test score information, relating to their students, was made available.

Although the practical effect on teacher expectations of prior or preconceived student achievement is an important factor in the consideration of professional judgement there is a need to

define and explore the nature and function of the underlying processes associated with this type of educational measurement. This view was tentatively indicated by Pedula, Airasian and Madaus (1980), commenting in particular on the effect of test results:

> "Little information exists on how actual test results relate to teachers' existing expectations, even though it is crucial to know this relationship in order to assess the potential for test results to affect expectations."
>
> (Pedula et al, 1980, p303)

The empirical work by Pedula et al (1980) concluded that the outcomes of teacher judgements reflected, in many respects, those obtained through standardized test utilization. However, the judgments encompassed behaviours beyond those identified by the standardized testing process alone. These academically related behaviours, for instance, attention span and persistence, illustrated the broader behaviour repertoire available for assessment purposes with the utilization of teacher judgement.

The need to find a theoretical framework on which to base the professional judgement process was most certainly advanced by further research work during the early 1980s. In particular, Borko and Cadwell (1982) adopted a comprehensive research design encompassing a variety of analytic techniques to investigate teacher decision strategies. Although the findings of this study were generally inconclusive, the research design and analysis aspects were adopted in part within later studies.

Cadwell and Jenkins (1986), for example, using a variation on the Borko and Cadwell design, probably provided the first evidence which could be used to support a theoretical framework and, therefore, a potential explanation of the judgement process:

> "The results suggest that teacher rating is a schema-based process in which the covariation among rating items is a function of teachers' implicit theories concerning the organization of student behaviours."
>
> (Cadwell and Jenkins, 1986, p460)

The actual framework proposed by Cadwell and Jenkins, for the rating (or judgement) process, involved two stages. The first was the formation, by the teacher, of a specific cognitive-model of the student under assessment. This model could be considered to be the product of numerous teacher-student interactions and teacher observations. The subsequent teacher rating is then arrived at by the unintentional comparison of this cognitive-model (representing the student) with the assessment criteria on which the rating is to be referenced. The term 'unintentional' is appropriate in this circumstance because, in reality, the teacher need only consider the specific and relevant evidence relating to the individual student characteristics which are to be judged. The various contributory aspects or features of the cognitive-model are, for practical purposes therefore, only required selectively and not necessarily collectively.

The construction and subsequent use of the student cognitive-model was critically analysed by Cadwell and Jenkins (1986). Three characteristics were identified within the framework as prerequisites which limited the function of the overall rating process. These were: Firstly, the student cognitive-model would inevitably contain errors, the magnitude of which are limited by the individual teacher's memory, perception and information processing abilities. Such errors are, however, compensated for or 'filled in' with information based on the content and context of the assessed behaviours. Secondly, the teacher may perform selective memory searches for behaviours consistent with the cognitive-model. This application reinforces the notion of self-fulfilment and negates schema inconsistent behaviour recognition or acceptance. Thirdly, the rating process itself contributes in a cumulative manner to the student cognitive-model. Each 'new' piece of information, obtained during the assessment process, is acreted to the evolving cognitive-model, thereby influencing the interpretation of subsequent assessment information. These concerns, in general, are summed up by Cadwell and Jenkins (1986) in the following comment:

> "These findings support the original claim that rating is a schema-based process constrained by the rater's information-processing abilities. Teachers were simply not able to rate one student characteristic independently of other information about the student."
>
> (Cadwell and Jenkins, 1986, p471)

Further research has substantiated Cadwell and Jenkins' schema-based theory of the rating process. In particular, Archer and McCarthy (1988), in a review of biases in student assessment, concluded:

> "Recent work in social cognition shows that under most circumstances behaviour consistent with pre-existing person schema is perceived more readily, and recalled more efficiently, than is behaviour which is incongruous with the schema; schemas influence our interpretations of ambiguous stimuli."
>
> (Archer and McCarthy, 1988, p144)

Additionally, Archer and McCarthy indicated that gender was generally a potential biasing factor and that this together with other possible biases could be eliminated, in part, by the adoption of 'blind-marking' whenever possible. This view of gender as a biasing factor is not universally supported; Borko and Cadwell (1982), for instance, reported no gender bias within their work. However, within the literature there is definitive support for the exclusion of gender as a biasing factor in the area of teacher expectancies of academic performance in particular. This conclusion is highlighted within the meta-analysis of 20 studies by Dusek and Joseph (1983) who commented:

> "This analysis leads to the conclusion that student gender is not a bias of teacher expectancies for general academic performance."
>
> (Dusek and Joseph, 1983, p331)

The dominant association of teacher expectancies of student academic performance with professional judgement, has probably retarded the development of a theoretical understanding of the rating process. In many respects the need to validate any theoretical framework for the rating process was precluded by the practical utilization of standardized tests as a 'benchmark' measure of judgemental worth. It is the all important aspect of judgemental worth which is the central theme of the next section. In particular, the validity of the theoretical schema-based process will be considered and its relationship with more conventional assessment measures explored.

3.3 Professional Judgement: issues of reliability and validity.

The comparability of professional judgement with the outcomes of more conventional forms of assessment is central to the argument for the acceptance of this technique within education. There is a significant body of evidence to support the contention that professional judgement can provide data which is not only valid but just as reliable as conventional assessment techniques (for e.g. Schroder and Crawford, 1970; Greenen and Smith, 1981; Greenen, 1984; Gullo and Ambrose, 1987; Wright and Wiese, 1988). For example, Wright and Weise comment on the broader assessment issue of experiential

relevance and familiarity and their effect on professional judgements:

> "The ability of these teachers to make accurate judgements based on their own experience and on their knowledge of important external measures suggests that experiential relevance and familiarity are important factors in any grading system that teachers will be able to use successfully."
>
> (Wright and Weise, 1988, p10)

The three principal and key studies cited in the previous section provided an indication to the reliability, validity and functioning of the rating process both practically and theoretically. To illustrate the significance of these studies each will now be considered in more detail.

Pedula, Airasian and Madaus (1980) provided firm evidence to support the notion that teacher judgment was comparable in terms of its reliability to standardized achievement tests. In a study involving 170 teachers and 2617 students, teachers' ratings of IQ, mathematics and English attainment were made and compared with standardized scores from IQ, mathematics and English tests together with additional ratings against 12 other social and academic behaviours. A subsequent factor analysis revealed correlations for the teacher ratings versus IQ, mathematics and English tests of 0.61, 0.63 and 0.65 respectively. These results provide support for the validity of the achievement ratings given by teachers when compared to standardized achievement tests.

Additionally these results illustrate an important consequence for the relationship between teacher expectancy and teacher ratings. It is possible that teachers' existing expectations of students apparently tap a dimension very similar to that of the corresponding standardized tests. The awareness of standardized test information for teachers, in these circumstances, merely serves to confirm and therefore reinforce their existing expectation of students' achievement potential. Hence, the accuracy of the rating process itself renders minimal any potential distortion of teacher expectation due to the prior knowledge of standardized test result information. The realization, however, that the rating process is more complex than may have been previously perceived was indicated more clearly through an inspection of the factor loading data.

Within the analysis, three factors were identified: one was essentially social behaviour based; the second was comprised academic classroom behaviours and teacher ratings on IQ, mathematics and English; and the third was comprised test scores in IQ, mathematics and English together with the corresponding teacher ratings. The teacher ratings loaded with the academic classroom behaviours (Factor 2) as highly as with standardized test scores (Factor 3). The significance and consequence of this is illustrated by the following comment:

..

"The results also indicate that teacher judgements of students' IQ, English and mathematics performance are confounded with their judgements of other academically related behaviours..... Not surprisingly, teachers cannot separate their judgements about academically related pupil behaviours which they observe on a daily basis from their judgements of pupils' standing on IQ, mathematics, and English."

(Pedula et al, 1980, p307)

Although this study provides initial and significant evidence for the reliability of professional judgement it tends to raise questions about the validity of this assessment technique. These two distinct aspects of reliability and validity will be the focus of the discussion of the second and third studies.

The findings of Hoge and Butcher (1984) provide further support for the reliability of teacher ratings as a viable alternative to standardized testing. A review of seven studies, including that of Pedula et al (1980), indicated an overall median correlation of 0.55 for teacher ratings of mathematics and English achievement against performance on corresponding standardized tests (although this is slightly less than the 0.63 to 0.65 range of the Pedula et al (1980) study for the equivalent measures). Their empirical work involved 12 teachers and 322 pupils. The teachers provided four rating measures including an achievement judgement expressed in terms of an estimated grade equivalent score for each pupil predicting the standardized achievement test score (subsequently administered). A regression analysis was performed with teacher achievement judgements as the criterion variable. The results

for the 298 eligible, complete data sets revealed that of the 12 teachers 10 showed the standardized achievement test variable constituted a statistically significant predictor of achievement ratings (six achieved significance at $p<.01$). IQ tests administered revealed scores which demonstrated overall significance as a predictor of achievement ratings, although only 3 of the 12 teachers individually achieved statistical significance. Finally, gender did not show overall or individual significance as a predictor of achievement ratings.

A further exploration of the biasing effects of IQ and gender was undertaken through an analysis of residual scores. These scores were formed from regression equations with the teacher achievement tests scores utilized as the criterion variable once again and the achievement test scores the predictor variable. With the exception of one teacher, statistical significance was neither reached individually or collectively for gender biasing. Similarly, IQ biasing, displayed no overall statistical significance, although 3 of the 12 did reach significance individually, with a further teacher very close to this ($p<.06$).

It is of importance that the teachers involved in this study were very experienced, of at least 6 years in all cases, and also that the achievement areas in question were relatively

well-defined. This prompted Hoge and Butcher to remark on the potential consequences of this fact:

> "It is possible that lower levels of accuracy would be obtained with less experienced teachers or with with other achievement areas."
> (Hoge and Butcher, 1984, p780)

In-spite of this consideration or limitation, the main conclusions reached from this study are both positive and supportive towards the reliability of teacher judgements:

> "The results of this study are on, the whole encouraging. They demonstrate that teachers are capable of making accurate judgements of the achievement levels of their pupils and that they are not overly influenced by pupil gender in making those judgements."
> (Hoge and Butcher, 1984, p781)

Finally, Cadwell and Jenkins (1986) returned to a feature first identified in the work by Pedula et al (1980), that of the validity of teacher judgements. Their empirical work involving the construction of an information-processing model of the rating process, described in a previous section, was tested by asking 18 teachers to rate 16 hypothetical student profiles. The profiles were formed by varying information along six different (profile) dimensions, including reading and mathematics achievement. The teachers were required to use each of the information profiles to complete a corresponding but not matching nine-item rating scale. The use of this item scale was intended as a means of allowing the raters 'policy' for integrating student characteristics to be 'captured'. In a typical 'policy-capturing' study, for instance Borko and

Cadwell (1982), teachers may be required to assess students' academic ability based on prior knowledge of standardized achievement test data - there is a direct link between the rating and the relevant available information. In this study, however, the use of relevant but not matching information was used, requiring teachers to infer student characteristics indirectly.

The data were analysed in two phases. The first, involved the use of regression equations, allowing the relative contribution of each profile dimension to be calculated. The second, utilized the technique of factor analysis, enabling a more detailed examination of the effects of the profile information to be ascertained. The initial analysis indicated that teachers attended to different profile information when rating the academic and non-academic items. The subsequent factor analysis of the nine-item rating scale data revealed two broad distinct underlying dimensions. These two dimensions indicated that teachers tended to distinguish between academic and non-academic behaviours - supportive of the regression analysis. The important finding from the study, though, was that the (statistical) removal of the student profile information from the analysis had little effect on the factor structure underlying the behavioural rating. The significance of this may be seen in the congruence coefficient values for the two factors, before and after the statistical removal of the

profile information was undertaken: 0.988 and 0.954 for the first and second factors respectively. Essentially, after the statistical removal of the profile information, any remaining correlation among the rating items must be due to the raters themselves. As Cadwell and Jenkins indicate:

> "Evidently, teachers imposed an organization on their ratings; that is, certain items were seen as "going together" and tended to be rated more similarly than they would have been had teachers based their ratings only on the profile information."
>
> (Cadwell and Jenkins, 1986, p470)

The value of this work is of particular importance when dealing with the validity aspect of teacher judgements. The recognition that the rating process is schema-based in which the covariation among ratings is a function of both known student behaviour and the teacher's implicit theory of student behaviour, needs to be considered carefully if acceptable levels validity are to be achieved using professional judgement.

The review of the three key studies illustrates several important points. The first identifies the reliability of the judgement process to be within acceptable limits in comparison with conventional assessments. However, the question of validity is unresolved. It appears the mechanism by which teacher judgements are undertaken involves confounding from other related behaviours or biases. These behaviours interact with the judgement process and produce distortions within the

rater's cognitive-model of the student under assessment. Hence, it is through the validity aspect of professional judgement that a researchable problem may be formulated.

The methods by which the issue of professional judgement validity could be investigated were evident within key aspects across all three studies. The ability to characterise and study the judgement process using regression and correlational analytic techniques was apparent within all three studies. The findings of the Cadwell and Jenkins (1986) study indicated the potential of the 'cognitive-modelling' with a 'policy-capturing' research design. Consequently, a variation on this research study and particular features of the analysis framework of the Borko and Cadwell (1981) study were developed for use in the research reported within this thesis.

3.4 Accountability: professional and managerial issues.

The professional notion of accountability within assessment generally centres upon the establishment of practices which ensure and promote acceptable levels of comparability between, and accuracy of, professional judgements. Inevitably such practices are focussed upon the performance of teachers and the rating mechanism itself. These aspects of professional, rather than managerial, quality control and quality assurance, are

more usually referred to as moderation and standardization (William, 1992a).

The growth in popularity and practice over recent years of SBA has seen a commensurate rise in the adoption and utilization of teacher judgments as an appropriate assessment technique. It has been argued that these judgments, when subject to appropriate moderation, can provide satisfactory assessment outcomes for external examination purposes (for e.g. Kingdom and Hartley, 1982). Additionally, Ingavarson (1990) has concluded from a study into the effects of a well established programme of 'consensus' moderation in Australia, that the process is very supportive of teachers and gives credibility to their assessments. The importance of moderation is further emphasized by Radnor (1991) who commented on the need for the moderation and assessment processes to be viewed in a cyclical manner. This cyclical description by Radnor of the assessment and moderation processes being mutually supportive and developmental does, in effect, delineate the function of standardization.

The gradual incorporation of the moderation process within that of standardization, applied to assessments administered by teachers, reflects in part the pressures from within schools for a greater degree of professional accountability. The need to promote cooperation within and across establishments is

paramount if comparable and acceptable teacher judgements are to be achieved within the broader context of SBA. The very subjective nature of the judgement process requires teachers to develop frameworks and strategies for the promotion of consensus between assessors and a greater understanding of the issues surrounding this assessment technique. However, there is an increasing pressure from Government and the Examination Groups for a greater degree of managerial accountability to be expected of assessments at all levels of education. Ironically, Pike and Murray (1991) argued that the desirable reduction of post-assessment moderation may be achieved by the use of performance indicators. Hence, managerial accountability, through the use of performance indicators, may well ultimately serve to promote a professional purpose; that of standardization.

This has been one worry of the teaching profession for some time. Gooding (1980) expressed concern over the use of external examination results as potential performance indicators. In particular this concern was directed to the use of performance indicators used to evaluate teachers rather than teaching. This possible inappropriate emphasis and over-reliance of external examination results as performance indicators is further illustrated by Bennett (1991), who indicated that pupil achievement was seen as a corollary by teachers for the legitimisation of their own professional standards. This

limited focus for performance indicators is not universally accepted though. Cuttance (1991), for instance, adopts a far more positive view and comprehensive approach to their use; indicating productive implications for various curriculum practices involving planning, development, delivery strategies, staff and management reviews.

3.5 INSET: a managerial policy to yield professional practice.

The professional demands for INSET in the area of assessment have witnessed greater attention more recently, initially due to the needs brought about by the introduction of SBA within the GCSE in 1987. This attention has continued with Government pressure for and the subsequent introduction of the National Curriculum in 1989, increasing the requests for assessment related INSET (HMI, 1991). Furthermore, there has also been the continual accumulation of evidence confirming that teacher participation in curriculum development provides for more effective innovation (Michael, 1987). In reality, INSET not only accommodates participation but its very nature actively encourages this. The increasing demands for INSET strategies to provide appropriate opportunities for teacher participation, within assessment at least, have themselves brought about changes within this provision and its subsequent evaluation. Assessment accountability is one of the themes central to the encouragement of educational change. In-Service Education

Training (INSET) is often the medium through which such change is managed. INSET is teacher centred and is essentially concerned with the professional development of the individual. However, as teachers form teams through departments, faculties, schools and LEAs, there are managerial issues which emerge. The need to affect change both consistently and effectively for the collective good of assessment practices specifically and education more generally is one such managerial problem.

The individual professional concerns of teachers have usually dominated the requests for INSET. Lawrence (1974), in a survey of in-service needs across 17 schools and 193 teachers (89 secondary), found that the nature of the part-time courses requested were predominantly related to teachers' individual everyday classroom practices. Few requests were made for courses which promoted whole school or broad curriculum development issues. This pragmatic, and in some respects insular, approach to INSET was probably due to the belief that key issues of classroom practice had not been addressed adequately by training programmes available up to that time. This possible neglect of certain areas of the school curriculum is detectable in the comments by Deale (1976) when reflecting on assessment issues:

> "Ideally assessment and evaluation should be kept in their proper place - that is treated as natural and essential components in any curriculum course. But, having stated the ideal, one must recognize reality too, in most cases teachers' knowledge of basic principles and techniques is so limited that one has to start with the ABC of assessment and to handle it properly would need a course on its own."
>
> (Deale, 1976, p206)

With such a vital area of the curriculum ignored either inadvertently or otherwise it is not surprising that teachers felt compelled at the time to demand INSET of what might be deemed a basic and a rather limited focus.

Any INSET provision, whether assessment related or otherwise, needs to be reviewed. Jasman (1987) has, for instance, looked at the development of in-service materials utilized within teacher training courses for development of teacher judgement techniques. The results of the work indicated areas of concern for this form of assessment. In particular, teacher judgments were found to be subject to a number of sources of invalidity. Similarly, the evaluation of the INSET materials and procedures were found to be inadequate in certain respects. This view, critical though it may be, highlights an important development in the evolution of INSET and its utilization in general. The need to view any in-service provision critically, through an appropriate method of evaluation, is crucial if the benefits of that provision are to be ascertained and its future use decided upon.

The effectiveness of INSET has not always matched expectations, often due to managerial factors. The Schools Council (1974) realized that a careful and considered approach was essential if aims were to be fulfilled through the pursuance of key initiatives or projects. The advocates of INSET though are to be found at all levels of education. Wheeler (1985), for example, has suggested several INSET strategies for the improvement of teacher training. These strategies covered a broad range of activities from greater cooperation between universities and schools to the encouragement of international projects. McGuiness (1985), however, concentrated more specifically on assessment within mathematics. He suggested the assessment model developed within his work could be, with modifications, generalized for other subjects.

These two examples illustrate the range of parameters within which INSET can function. At one end of the spectrum training exists as a series of strategies, with large scale managerial difficulties. At the other end of the spectrum the training needs are more subtle involving the localized dissemination of knowledge and skills with small scale managerial problems. Whatever the scope of the INSET initiative, small or large scale, the key to effective innovation probably lies in the mobilization of teachers (Ilsley, 1989) and this in turn has professional implications. It is the professional development

feature of INSET which may be seen to ultimately determines the likelihood of success of any educational initiative.

In-spite of certain managerial difficulties associated with INSET (for instance financial constraints) there still remains the consensus of opinion that its outcomes, under favourable circumstances, are professionally beneficial. It is the reconciliation of the managerial and professional demands of INSET which are very much the focus of concern for teachers and educational planners alike. With many training priorities and limited budgets and resources it is inevitable that tensions will always exist between the managerial and professional needs of INSET. Troman (1989) argues that the tensions between the managerial and professional demands of centralized and school-based-assessment may become intolerable. Ultimately, the conclusion to this argument is that of choice between the two competing assessment systems. Unfortunately, this is not helped when, as Hannan (1985) points out, the motivations for the adoption of certain assessment techniques is sometimes for political reasons of convenience rather than utility and reliability.

3.6 Discussion.

The aims associated with teacher judgments are essentially professional, although there are a few which are managerial. The professional requirements of teacher judgments are, however, quite considerable. The operation of a frequently highly subjective process within an overall SBA framework is the assessment reality which is of specific interest and concern to the teaching profession. Moreover, the requirement of teacher judgments to contribute positively and with effect to SBA procedures is not only desirable but essential for the long term future of this form of assessment. In contrast to teacher judgement, there are significant direct managerial implications for SBA in general as a credible practice within external examinations (p31). These are associated with the promotion of consensus and consistency within an SBA framework incorporating professional judgement.

In-spite of the reservations directed towards significant teacher involvement in student assessment through the rating process (for example, Torrence, 1988), the past thirty years have seen a significant body of evidence accumulate to quell such concerns. Schroder and Crawford (1970), for example, indicated that teacher judgements of academic achievement are an important and dependable adjunct to the use of standardized achievement tests. However, to qualify the term 'dependable'

requires the consideration of reliability and validity estimates. There is some consensus regarding professional judgment in terms of both attributes, although certain researchers have questioned the validity of the rating (or judgement) process itself. More specifically, the function of this process has been found to possess a degree of complexity unexplored before the 1980s. Recognition of this complexity is illustrated within several (educational) reports. Although a review of the literature shows these not to be in abundance, there were sufficient studies to enable the delineation of a researchable problem to be undertaken. Additionally, it was also possible to develop a research design and analytic framework capable of exploring this problem.

Broader issues of experiential relevance and familiarity of the subject under assessment by the teacher have been shown to be pertinent to the judgment process. The terms relevance and familiarity, in this context, embody two of the essential prerequisite managerial and professional conditions for the successful use of teacher judgements. Hence, it is arguable that the managerial and professional aims of teacher judgment are unlikely to be fulfilled without adequate attention afforded these two features. In reality, the objectives by which these aims could be fulfilled are assessment accountability and INSET.

The notion of accountability within assessment is frequently viewed as a political call reflecting the desire to produce evidence on the effectiveness on teaching - with the teacher often as the focus (Eggleston, 1979). Although accountability in this situation has both managerial and professional features, it is usually the managerial use of performance indicators which tend to be the focus of attention. The consequences of this managerial emphasis on the evaluation of teaching is thought by some to be to the detriment of the professional development of teachers (Erskine, 1987). It is this professional aspect of accountability which is serviced predominantly by INSET. The introduction and subsequent utilization of SBA, with professional judgement as its principal assessment technique inevitably, demands a great deal of the INSET process. However, failing to confront the issues and manage carefully the INSET needs of assessment accordingly can itself cause problems.

The consequences of indecision are exemplified by the comments of Nelson (1988) who expressed concern over the anxiety and lack of expertise directed towards the introduction of GCSE mathematics coursework and the requisite assessment and teaching strategies necessary for its implementation. There was clearly a major role and demand for the utilization of INSET. However, as William (1992b) indicated, the use of development

and evaluation strategies employed by educators towards INSET had been questionable at best, up to that time:

> "We were aware that many evaluations of inservice provision conducted in the past had centred around the participants' reactions to the training, recorded at the end of the course and amounted to little more than asking teachers 'did you have a nice day?'"
>
> (William, 1992b, p8)

The important and major role for INSET within the development of teachers is of some consequence for the promotion of professional judgment as a reliable and valid assessment technique. The need for teachers to view the judgement process within the confines of a 'relevant' assessment arena and have the 'familiarity' with the technique to give confident and accurate measures is the key to the future success and indeed continued existence of SBA.

3.6.1 Summary.

Professional judgement as a legitimate internal examination practice became accepted during the latter part of the 1980s. Its ability to allow the assessment of curriculum aspects hitherto unassessable has been a significant step forward within the realms of educational measurement. However, the acceptance of teacher judgment as a reliable and valid assessment technique is not shared by all. Educational managers may view this technique as a possible threat to standards. Some researchers have found a complexity within the judgement (or

rating) process which poses technical questions about the assessment and its 'true' validity. It is through the process of accountability and staff development (INSET) that a solution to this validity problem resides. The need to view within context the development of professional judgement is essential if this solution is to have meaning. Hence, the next chapter will review the initiatives leading up to and including the introduction of the National Curriculum. This will provide the necessary historical and contextual background within which professional judgement operates.

Chapter 4.

A National Curriculum: from theory through policy and into practice.

The aims of this chapter are to outline some of the main driving forces behind curriculum and assessment development during the mid to late eighties. It will highlight the introduction of a national assessment framework (the National Curriculum). Finally, teacher judgement will be placed within the context of this national assessment framework; and some difficulties associated with its use explored.

4.1 Introduction.

The realization of the need for major assessment change was probably first highlighted within the Cockcroft Report in 1982. Although this report was directed at Mathematics, there were significant implications for other curricular areas in particular regarding the importance of teacher assessment. The mid 1980s saw a period of sustained critical comment, regarding assessment practices, from various quarters. Goldstein and Nuttall (1985) expressed concern over several problems associated with external examination system utilized at that time. This view was, in many respects, typical; citing issues of a curricular and assessment nature. Criticism was not restricted exclusively to the domestic scene. In the United

States disillusionment found support with the advocates of a centralised curriculum and assessment framework. In this respect the National Assessment of Educational Progress (NAEP) project was considered by Ferrara and Thornton (1986) to be a candidate for a possible national achievement test.

Within the UK, it is probably for more pragmatic reasons that major change was ultimately initiated during the latter part of the 1980s. This decade in particular witnessed substantial demographic, social and economic changes. Educational development and growth reflected these changes through key statutory initiatives such as the TVEI and the GCSE (Barnes, 1987). Inevitably, particular initiatives can be identified in history as being central to the evolution of assessment. The work of the Assessment and Performance Unit (APU) and the Graded Assessment (GA) movement provided the main impetus for assessment change during the 1980s. These are reviewed in the next two sections.

4.2 The APU: national assessment in theory.

The primary purpose of the APU was to research into various educational assessment aspects and to report results (Black, 1984). The many reports provided a source of evaluation data on key assessment and curricular activities which has been of interest to many within education. The main theme of many of

the projects undertaken by the APU was concerned with assessment and its effect on teaching styles. Several reviews have focussed on this aspect of the APU's project work, for example: Preston (1980), Broadfoot (1980) and Foxman and Mitchell (1983).

Within education there is a strong belief that current assessment techniques need to evolve, enhancing their curricular validity. The review work, undertaken by many researchers, on the various APU evaluation data has reflected this notion. Bell (1977) highlighted the importance that assessment should cover a range of student outcomes; not just content - process and attitude for instance. Similarly, Stones (1979) indicated the technical short-comings of more traditional testing techniques and looked at available assessment alternatives. Murphy (1988) also concluded that a move from traditional learning strategies to more student centred ones was required for the future. There is little doubt that the APU has provided evidence for the evaluation of many assessment practices. However, it is the contribution of the APU to national achievement testing which is probably its most noteworthy, and certainly its most controversial.

The notion of national assessment was considered in some detail by the APU during the latter part of the 1970s. Driver and Worsley (1979) described particular national methods of

assessment and the monitoring of achievement in science. Key tasks were envisaged for 11, 13 and 16 year olds under this framework. Earlier, Marjoram (1978) considered the potential of a national assessment system and its possible use for improving the student transfer process between schools. A cautionary note was made by Galton (1979), however, comparing the APU's assessment strategies with those of the NEAP project in the United States. This concern was reflected by Gooding (1980) who surveyed the views of 124 teachers (within the UK) on national achievement testing - the findings indicated a strong opposition to this style of centralised assessment.

In conclusion, the APU provided a substantial body of evidence for the limitations of traditional assessment techniques. The plethora of evaluation data enabled researchers to develop strategies for the promotion of more effective assessment practices. The APU concentrated specifically on the general development of national achievement testing strategies. This work, over a period of years and through several large scale initiatives, made the notion of national assessment a theoretical possibility. Although the APU provided the means for a national assessment scheme its practical reality involved curricular considerations. In this respect the Graded Assessment movement was instrumental in the conversion of national assessment from a theoretical possibility to that of a practical reality.

4.3 Graded Assessment: national assessment in practice.

Graded Assessment as a notion has been a topic of discussion for many within the educational literature (for e.g. Cockroft, 1982, Pennycuick and Murphy, 1986, Gipps, 1990). From a review of this literature two key features emerge which characterise Graded Assessment schemes. Firstly, these schemes are modular in form (Nuttall and Goldstein, 1984). Each scheme is sub-divided into units or topics of work each with specific objectives. This 'goal-orientated' approach is undertaken through the utilization of module assessments - targeting the objectives of the material covered within the unit or topic. These module assessments may involve the use of time-limited tests; teacher ratings or a combination of both. The second characterisation is that assessments are allocated to particular levels - these levels may or may not form prerequisite hierarchies.

At least five Graded Assessment schemes have been developed and successfully implemented within secondary schools during the latter half of the 1980s. Each scheme has been responsible for the enhancement, and sometimes development, of particular curricular features. Swain (1991) has indicated the role of the Graded Assessment in Science Project (GASP) in the promotion of scientific explorations. Within modern languages Page and

Hewett (1987) indicated how Graded Assessment has contributed to the overall development of the subject. Specifically, the Graded Objectives in Modern Languages (GOML) approach has significantly influenced the pedagogy associated with the teaching of this subject. Of the five schemes available it is probably that of Graded Assessment in Mathematics (GAIM) which has utilized to greatest effect the notion of learning hierarchies. As Brown (1989) has pointed out there is substantial evidence for the success of this Graded Assessment scheme, motivating and promoting greater mathematical achievement.

Graded Assessment, however, is not without its critics. Acknowledgement was made by Pennycuick (1987) of the curricular and administrative difficulties associated with GA schemes. Noss et al (1989) provided substantial criticism of the GAIM scheme; citing the inappropriate use of learning hierarchies as a particular concern. There is little doubt that the organizational features of Graded Assessment schemes have administrative burdens beyond those associated with conventional assessment procedures at GCSE. However, the uptake of GA by 57,000 students in over 300 schools has provided a significant argument for its practical workability (Portal, 1991). Similarly, the use of learning hierarchies within GAIM may simply present an 'effective' practical framework which is empirically based. Certainly the learning hierarchies on which

GAIM is based were developed empirically (Hart, 1979) - thus lending support to the notion of an 'effective' and 'valid' framework.

Graded Assessment illustrated the first real attempt at integrating learning and assessment strategies within schemes designed for external examination certification. The adoption of unconventional assessment techniques has prompted the evaluation of subject syllabus design, teaching methodology, resource utilization and the consideration of research issues. Finally, a 'small scale' working prototype for national achievement testing was realized through the linking of certain Graded Assessment schemes to the GCSE. The subsequent development and implementation of a 'large scale' working model for national assessment will be detailed in the next section.

4.4 The National Curriculum: from policy to practice.

The passing of the Education Reform Act (ERA) in 1988 created the beginnings of a basic curriculum linked to a nationally defined assessment framework; both of which were statutory and therefore compulsory although this did not apply to private schools. Whilst the curriculum content on a general and subject specific level was dealt with by the National Curriculum Council (NCC), the national assessment framework was within the purview of the Task Group on Assessment and Testing (TGAT).

TGAT's single purpose was to produce recommendations on which the National Curriculum for England and Wales could be established.

The deliberations of the TGAT resulted in the production of a report which contained recommendations for the implementation of a national assessment system. The TGAT report contained several elements, three of which were fundamental and probably embody the underlying philosophy of the proposals overall: The first concerned itself with the construction of an assessment system capable of meeting key criteria of being progressive, able to moderate, formative and utilise criterion-referencing. In accepting that previous systems had not accommodated these features they comment:

> "Our task has therefore been to seek to devise such system afresh. We believe that the model of assessment put forward in this report builds on some existing good practice and represents an advance on assessment practices in other countries.
> (TGAT, 1988, para. 13)

Further to this, acknowledgement was made of the opposition to national assessment voiced within many educational quarters:

> "But we could not approach this task without also recognizing that many are deeply opposed to any system of national assessment and testing"
> (TGAT, 1988, para. 13)

The second fundamental feature of the TGAT report focussed on the use of a unique attainment grading system. This grading system was designed to convey effectively the attainment of students based primarily on a series of criteria or Statements

of Attainment. For curricular purposes these criteria were organized within related areas or Attainment Targets. For reporting purposes Attainment Targets were clustered within Profile Components. It was intended that the aggregation of Attainment Target scores within a Profile Component would provide the grading mechanism to be used within the envisaged reporting procedures. The essential elements of this grading system were described as follows:

> "We recommend that each of the subject working groups define a sequence of levels in each of its profile components, related to broad criteria for progression in that component. For a profile component which applies over the full age range 7 to 16, there should be ten such levels, with corresponding reduction for profile components which apply over a smaller span of school years."
>
> (TGAT, 1988, para. 101)

The third and final statement of intent within the TGAT report documented within the report focussed upon the mechanics of the assessment process. Acknowledgement was given that previous good practice and key advancements in assessment procedures should be accommodated within any proposed scheme. Consequently, the notion of traditional time-limited tests were re-conceptualized in terms of Standardized Assessment Tasks. Further to this, assessments (or professional judgements) undertaken by teachers within the context of the normal classroom experience were to be given a substantive role within the overall assessment process. This proposed partnership between formal and informal assessment is highlighted within the following recommendation:

> "We therefore recommend that the national assessment system is based on a combination of moderated teachers' ratings and standardised assessment tasks."
>
> (TGAT, 1988, para. 63)

The importance attached to the professional judgements of teachers was a prominent feature within the report. Consequently, TGAT viewed the proposed assessment framework as unique in its construction, and progressive in its envisaged operation. However, this sentiment was not universal. Gipps and Goldstein (1989) concluded, after careful scrutiny, that the report's recommendations were not as progressive as may have been first thought. Similarly, Allanson et al (1990) questioned the rationale behind the assessment framework - with reference to two aspects in particular. The first concerned the level of meaning behind the reporting of attainment through the use of Profile Components. Their view was that the process of aggregation across Attainment Targets could give a misleading impression about a student's actual achievement. The second concern was associated with the degree of public confidence in the results of assessments undertaken by teachers. The essence of their argument centred upon the notion of comparability between the ratings (or judgements) of different teachers. The adoption within Scotland (1986) of a criterion-related assessment framework for the Standard grade examination indicated the difficulties associated with the comparability of ratings across different teachers. In particular, the concern over public confidence in the results of teacher assessments is

indicated within Sharpe's comment regarding the use of Grade Related Criteria (GRC):

> "This modified role for GRC has not been the result of any diminution in the importance which is attached to the aim which they represent; rather it has resulted from practical experience which demonstrated the limitations of criterion referencing within the context of summative assessment for public certification."
> (Sharpe, 1991, p16)

Nuttall (1992) questioned the overall validity of an assessment process whose development was not linked to curriculum practices and the lack of provision for achieving these. Thomas (1989), for instance, indicated the training implications for the introduction of the National Curriculum. A failure to meet these demands allied to the subsequent increased professionalism expected of teachers (Chard, 1990) became of sufficient contention to merit concern from teachers. However, Osborne (1991) speculated that in-spite of the difficulties associated with the introduction of the National Curriculum, teachers may well ultimately internalize the changes achieving ownership of the educational reforms.

It is the issue raised by Nuttall regarding the separation of curriculum and assessment aspects which reflected a concern relevant to the whole development process associated with the implementation of the National Curriculum. The fact that curriculum and assessment were viewed as distinct entities was the cause of some anxiety (Atkinson, 1990). The creation of the

Schools Examination and Assessment Council (SEAC) in 1989, as a replacement for the SEC, confirmed the significant division between curricular and assessment issues. In this circumstance, it is understandable that the implementation of the National Curriculum was problematic. The active policy pursued by government of separating the prominent features of curriculum and assessment ensured the continuation of the problem.

These difficulties, variously cited, were viewed by the Secretary of State for Education, at that time John MacGregor, to be challenges to the teaching profession. In particular, he was seen to counter criticism directed towards the Statements of Attainment and the ten level scale within the National Curriculum with the following comment:

> "After all, the gradations in the 10-point scale are very broad ones, and the criteria defining them are pretty clear."
>
> (DES, 1990, p14)

However, in the absence of specific evidence to substantiate this statement it was unlikely to quell the disquiet within the teaching profession directed towards the use of statements of attainment.

It is difficult to discover literature which is wholly supportive of the implementation of the National Curriculum. For instance, Jarman (1990) pointed to the potential of the National Curriculum for aspects of cross-phase continuity, although this was tempered with considerations of its several

limitations. Similarly the HMI (1989) reviewed the progress of the curriculum's implementation within 500 schools and concluded that curricular issues were been positively and successfully addressed. However, the assessment and recording aspects of student attainment were causing uncertainty and anxiety. In a follow-up survey of 100 schools from the original cohort, the HMI (1990) found anxiety towards assessment issues remained. A further report by the HMI (1991) highlighted the nature of this anxiety more specifically. The two key issues of concern were the meaning of 'mastery' and its use with Statements of Attainment; and the achievement of consistent standards when using these criteria.

The questionable attributes linked to the assessment procedures within the National Curriculum were brought into focus by many, even before the process of implementation had started. Nuttall (1988) indicated the existence of certain unresolved psychometric issues. Hartnett and Naish (1990) considered the limitations of the initial consultation document utilized within the early years of implementing the National Curriculum. In particular this criticism centred upon the tenuous relationship between documented solutions and the corresponding problems. A broader perspective was adopted by Longstaff (1990) questioning the compatibility of the National Curriculum with the fundamental principle of democracy. Finally, Broadfoot (1991), in a survey of 88 primary teachers, indicated that

there was little support for the role of Standardised Assessment Tasks; which were not viewed as positively contributing to the assessment process.

It would be inappropriate to criticise the National Curriculum in its entirety when the majority of views indicate the assessment procedures to be at fault. As Raban (1991) reported, most teachers have welcomed the impetus the National Curriculum has given to reflect critically on their own practices. Although assessment problems associated with 'busy' classrooms were also cited as an added pressure for primary teachers. With reference to the technical problems of classroom assessment, Gipps (1992) criticises the complexity of the assessment structure within the National Curriculum.

It is the notion of levels within the overall assessment structure which has been a controversial issue since its recommended adoption by TGAT. Lofty (1990) illustrated the difficulties of a hierarchical curriculum within the United States and placed this in the context of the British National Curriculum. Level development was also presented as a problem by Relf (1990) when considering the diverse levels of content within the Mathematics National Curriculum. A pragmatic issue of level achievement was highlighted by Shayer (1991), indicating a significant shortfall in the expected number of

students gaining higher levels within the Science National Curriculum.

The very specific problems associated with the notion of levels are one example of the practical concerns that teachers faced when implementing the National Curriculum. The difficulties associated with the assessment aspect included issues of: the compatibility between diagnostic and summative purposes (Mortimore, TES 12th July, 1990, p12) and the validity of assessing against individual Statements of Attainment (SMP, 1990). These issues exemplify the mismatch between the theoretical aspects of assessment and their implementation. In particular, this is probably of greatest concern when teachers undertake their own assessments or professional judgements.

4.5 Discussion.

Part of the background to the introduction of the National Curriculum in 1989 was a perceived need for change within assessment practices. The plethora of research questioning the validity and reliability of previous practices became a strong argument for the implementation of a new style of assessment. The TGAT report provided a framework on which national assessment could have been based. Two working models, in particular, were available for scrutiny; and which could be used to evaluate the framework. The Graded Assessment model

provided a workable and reliable scheme (Portal, 1991) through which the TGAT proposals could potentially be fulfilled. Indeed Black (TES, 14th July, 1989, p11) considered the Graded Assessment schemes to be the closest to the TGAT model. In contrast the Standard Grade examination model introduced in Scotland in 1986 provided an illustration of the potential problems associated with the use of a criterion-referenced assessment scheme on a national scale. (Sharp, 1991).

It was the aspect of building on existing 'good' practice that the TGAT believed to be of some importance for the success of the assessment model. Yet, the assessment model adopted within the National Curriculum paid little attention to the examples of 'good' and 'bad' practice depicted within the Graded Assessment schemes and the Scottish Standard Grade examinations respectively. The consequences of disregarding this point were possibly reflected in the number of articles critical of the National Curriculum in the early years.

In conclusion, the National Curriculum presented a problem, centred upon the assessment procedures proposed by TGAT and implemented by SEAC. The level structure, advocated for each subject, presented some difficulties - the levels forming a continuum on which criteria were referenced. The specific concern of teachers, however, was related to the 'clarity' of the Statements of Attainment utilized as criteria within the

assessment scheme (SMP, 1990). The credibility of any process of professional judgement is limited by the quality of the criteria on which teacher judgements are based. Hence, within the National Curriculum the utility of such judgments referenced to Statements of Attainment was brought into some degree of question (Griffiths, TES, 5th February, 1993, p2). However, the credibility of the judgement process is limited also by the quality of the judgement policies (or decision strategies) on which these are based (Wakefield, 1980). Concern over this particular aspect of professional judgement is not evident within the literature.

4.5.1 Summary.

The importance of professional judgment, within the National Curriculum, brought into focus the assessments undertaken by teachers with a degree of scrutiny hitherto uncalled for. The judgement criteria have seemingly been the focus of much attention within the implementation phase of the National Curriculum. However, the criticism that Statements of Attainment are ambiguous or lack clarity, and are therefore unsuitable criteria on which to base ratings, has not been equalled by questioning of the adequacy of the judgement policies adopted by teachers. Probably a more appropriate focus should be the interaction between the teacher's judgement policy and the criteria to which this is applied. The next

chapter will consider this interaction as the basis of a researchable problem in more detail and outline the initial phase of the empirical work undertaken within this study.

Chapter 5.

The Research Problem: a preliminary investigation of the issues.

This chapter will outline the development of some specific questions associated with Teacher Assessment into a researchable problem. It will describe the use of a preliminary interview schedule and small-scale questionnaire survey of a sample of secondary school mathematics teachers. The findings of these two investigations will be related to the review of literature. Finally, the chosen aims of the study will be detailed and described through a series of testable hypotheses.

5.1 Introduction.

The review of literature (chapter 2) indicated the increasing importance of criterion-referencing within the context of School Based Assessment. This form of assessment gained in popularity due to its ability to provide measures which are considered to have greater validity than conventional means, time-limited testing for instance. The fulfilment of curriculum validity together with a degree of reliability comparable to conventional assessment provided CRM (within a SBA framework) sufficient credibility to allow its continued utilization and further development within education.

The adoption of assessment techniques which rely upon teachers undertaking a key role highlighted professional judgement as an area or issue of concern (chapter 3). A professional judgement was considered to be the interaction of the teacher's decision strategy applied to a criterion. The review of literature identified this interaction as a complex process, with several inter-related aspects (p37). Although it was found that valid and reliable professional judgements were possible, certain pre-requisites need to be fulfilled. The first involves the development of teachers' assessment skills, essential if professional judgements are to be effective. The inevitable use of INSET to enable the delivery of such development was highlighted within the literature as problematic (p54) in this respect. INSET provision has to fulfil both professional and managerial requirements. The effective development of teachers' decision strategies, which would require INSET with a professional focus, are unlikely to be met because of managerial constraints. The second requires the adoption of well defined (and specified) judgement criteria (p18). These criteria may be seen as a limiting factor of the professional judgement process. If a criterion is ill-defined it may be impossible to achieve an effective judgement irrespective of the appropriateness of the decision strategy adopted.

The review of the literature relating to the implementation of the National Curriculum (chapter 4) revealed a large degree of

criticism directed towards the criteria developed for use within the associated assessment procedures (p77). Concern regarding the ability of teachers to undertake effective professional judgements focussed upon the Statements of Attainment (assessment criteria). The issue of teacher decision strategies as an essential element of professional judgement is not documented to any great extent within the contemporary National Curriculum literature. Similarly, the demands for INSET were made in order to accommodate and compensate for the perceived deficiencies in the Statements of Attainment rather than any problems associated with teacher decision strategies.

In the context of earlier research into teacher decision strategies possibly a more appropriate question which remains unanswered within the contemporary National Curriculum literature is: "To what extent are inadequacies associated with the 'professional judgements' of teachers due to the application of inappropriate decision strategies rather than the utilization of deficient assessment criteria?" The answer to this question is of some importance if the utility of Teacher Assessment within the National Curriculum is to be established.

5.2 The Preliminary Investigation: method of administration.

The preliminary investigation was undertaken during the latter half of 1990. Although the issue of professional judgement within the National Curriculum is generic (affecting all subject areas), the delineation of a researchable problem required a specific focus of attention. Previous research by the author (Atkinson, 1990) relating to the introduction of the Mathematics National Curriculum at Key Stages 3 and 4 provided essential background information within this subject area. Similarly, the choice of researchable subjects was essentially limited by the phased introduction of the National Curriculium. Mathematics or Science, specifically at Key Stage 3, provided the only practical curriculum areas for this investigation. Hence, the focus of attention for the researchable problem was the Mathematics National Curriculum at Key Stage 3. This work is described in three parts: administrative procedure; sample; results and findings.

5.2.1 Procedure

The initial interview schedule was completed at four Humberside secondary schools involving a total of seven mathematics teachers during one week in November 1990. The teachers ranged in responsibility from Heads of Faculty to Main Professional Grade teachers. During the interviews each teacher was asked to

comment upon issues concerning four separate aspects regarding the implementation of the Mathematics National Curriculum. These aspects were: Statements of Attainment utilized as rating criteria; rating criteria exemplars; rating criteria within the context of the 10 Level scale and Attainment Targets; and teacher pre-conceptions or biases regarding student attainment. The use of prompting was restricted to a degree sufficient to elicit a response of sufficient length or detail to make clear the respondents' feelings or experience. A summary of the findings from the interview schedule are depicted in Table 5.1.

Table 5.1. Concerns of Teachers recorded during the interview schedule.

Curriculum aspect/feature	Total mentions
Rating Criteria (or SoAs):	
vary in quality;	4
multiple-interpretation possible;	4
Rating Criteria exemplars:	
examples vary in quality;	3
examples can be over prescriptive;	3
Rating Criteria Levels (and ATs):	
levels can be difficult to conceptuálise;	3
Teacher Pre-conceptions/Biases:	
preconceived ideas may influence the assessment process;	4
teacher preconceptions no problem;	3
consistent interpretation would be achievable with time;	4

The responses obtained during the interview schedule have been grouped within the four aspects. It was noticeable that when

asked to comment within any one category *all* the teachers interviewed mentioned one or more of the other curriculum aspects or features included within the interview. There appeared to be a single underlying concern within all of the four categories of questioning. A non-committal approach was adopted by one teacher questioned regarding the notion of rating criteria in the context of Levels (and Attainment Targets). Such a cautious approach was justified on the grounds of unfamiliarity with Teacher Assessment at the time of questioning. However, the responses generally confirmed some of the key concerns expressed within the literature regarding the difficulties surrounding the issue of Teacher Assessment within the Mathematics National Curriculum.

A questionnaire was drawn up primarily from the interview schedule responses. Although similar to the original interview schedule categories, some revisions were made to produce the questionnaire section headings. These revisions included an additional section relating to a specific issue: the notion of stability in a student's assessment performance. The review of literature indicates the importance of this issue through the Concepts of Secondary Mathematics and Science study (Hart, 1980) and the Graded Assessment In Mathematics initiative (GAIM, 1988). Both highlight concerns with the specific consideration of 'short-term-retention' effects on students' assessment performance. Because of its importance, it was

expected that the issue short-term-retention would have featured as a response within one or more of the categories of questioning collected during the interview schedule. Finally, with the semi-structured open response format of the questionnaire it was intended that only the key or main concerns would be reported within the findings. A main concern was defined as one mentioned by five or more respondents.

To encourage a high return rate the questionnaire was limited to two sides of a single A4 sheet and contained only six sections (Appendix 1). The first requested biographic details including number of years teaching and experience of teacher assessment. The second section focussed upon the issue of Statements of Attainment and asked for comments, both positive and negative, regarding the appropriateness of Statements of Attainment as assessment criteria. The third section dealt with the issues of teacher pre-conceptions or biases affecting student attainment. Bias exemplar categories were included; for example: gender and presentation or neatness of work. Section four concerned the stability of student assessment performance. This aspect was allocated the generic name of 'sustainability' and attempted to isolate any effects within the contexts of the knowledge, skill or understanding attributable to the student. The fifth section centred upon the contextual features of examples and levels. Comparability and utility issues formed the basis of this area of questioning. The sixth and final

section asked for views on the most 'important issue' confronting Teacher Assessment within the Mathematics National Curriculum.

5.2.2 Sample.

Several copies of the questionnaire, together with covering letters, were sent to the four Humberside secondary school mathematics faculties originally involved with the interview schedule. Frequent contact between the heads of faculty within each school and the author pre-empted the employment of follow-up procedures. During a three month period between December 1990 and March 1991 replies were received from 14 of the 25 potential respondents, a return rate of 56%. The biographic data is summarised in Table 5.2.

The length of service of the questionnaire sample ranged from 6 to 25 years (M=14yrs and SD=5.7yrs). In this respect, the sample of respondents illustrated would appear to represent a group of teachers with a significant familiarity of teaching Mathematics and assessment. Their substantial experience would probably imply an awareness of the relevant issues regarding Teacher Assessment within the National Curriculum. Hence, although small, the sample of respondents would appear to be a suitable group of teachers to represent informed opinion on the

issue of Teacher Assessment within the Mathematics National Curriculum.

Table 5.2. Teachers' biographic information.

Service Information		Number
Length of Service	0 - 4 yrs	0
	5 - 9 yrs	4
	10 - 14 yrs	6
	15 - 19 yrs	2
	20+ yrs	2
Principal Subject	Mathematics	14
	Non-specialist	0
Assessment Experience		**Number**
Criterion-Referencing	Yes	11
	No	3
Teacher Assessment	Yes	12
	No	2

5.2.3 Results: summary of responses.

In Table 5.3 the responses obtained have been grouped within the other five sections of the questionnaire. These are Statements of Attainment; pre-conceptions; sustainability; levels and examples; and important issues. The responses to sections two through six are illustrated in Table 5.3.

The general concerns, expressed within the second section of the questionnaire responses, regarding the utility of

Statements of Attainment as rating criteria, appear to be consistent with the published literature. The criticism of being too 'broad' or 'general' reiterates the comments of many related articles reporting research upon assessment within the National Curriculum (Gipps, 1990).

Table 5.3. Concerns documented within the Teacher Assessment questionnaire.

Curriculum aspect/feature	Total mentions
Section 2 - Statements of Attainment:	
criteria too broad and general;	9
criteria of multiple meaning;	5
Section 3 - pre-conceptions:	
pre-conceptions are not a problem;	6
any are compensated for by the teacher;	5
Section 4 - sustainability:	
a one-off demonstration is acceptable as evidence of sustainable attainment;	5
sustainable defines a performance which is repeatable over a long period time;	6
Section 5 - levels and examples:	
Comparison between levels difficult;	7
Prescriptive nature of examples unhelpful;	6
Prescriptive nature of examples helpful;	7
Section 6 - important issues:	
Curriculum evaluation and monitoring of assessment standards;	12

The issue of teachers' pre-conceptions of student attainment, identified in section three of the questionnaire, was considered by the majority of respondents to present few if any

difficulties. This view is not altogether supported within the literature. Teacher expectations are considered to be a key factor influencing perceived student performance (Cooper et al, 1982). The fact that teachers may be unaware of such influence is also a feature of the literature (Shalverson and Stern, 1981). For this reason the notion of teacher pre-conceptions of student attainment may not be readily dismissed.

Within the fourth section of the questionnaire, the responses indicate two interpretations attributed the term sustainability. The first asserts that a single demonstration of attainment is adequate for the permanent acquisition of the criterion. The second interpretation stipulates that any demonstration of attainment be repeatable at a future time to confirm the permanence of the acquisition. Although each view was supported by approximately half of those questioned, 11 respondents mentioned the need to consider 'temporal' effects within any discussion of sustainability in agreement with the published literature. Sustainability, as defined by the Graded Assessment In Mathematics development group, for instance, require such demonstrations to be durable beyond a notional two-week limit (i.e. the assessment is conducted two-weeks or more after any related teaching).

Responses regarding 'levels and examples', covered by section five, illustrate a difference of opinion within the

respondents. The use of judgement criteria exemplars elicited either a negative or positive comment by 13 of the 14 respondents. The responses were fairly evenly balanced between the advocates and non-advocates of the prescriptive nature of the judgement criteria. There is a similar degree of dissagreement over the difficulties teachers confront in the process of comparing levels across Attainment Targets.

Finally, 'important issues' which constituted the sixth section of the questionnaire elicited numerous opinions and concerns. Although this is a generic title and includes a broad range of issues. A total of 12 out of the 14 respondents made either direct or indirect reference to collective issue of assessment evaluation and monitoring. The cited concerns over issues of assessment evaluation and monitoring reflect general the stand-point of the published literature.

5.3 Discussion and Comment: formulating aims and hypotheses.

Teachers' responses to the interview schedule and questionnaire survey provide a consistent description of the perceived difficulties surrounding the use of Teacher Assessment within the Mathematics National Curriculum. These perceptions have substantial support from the contemporary National Curriculum literature. However, they are not entirely

consistent with the more established literature within this field.

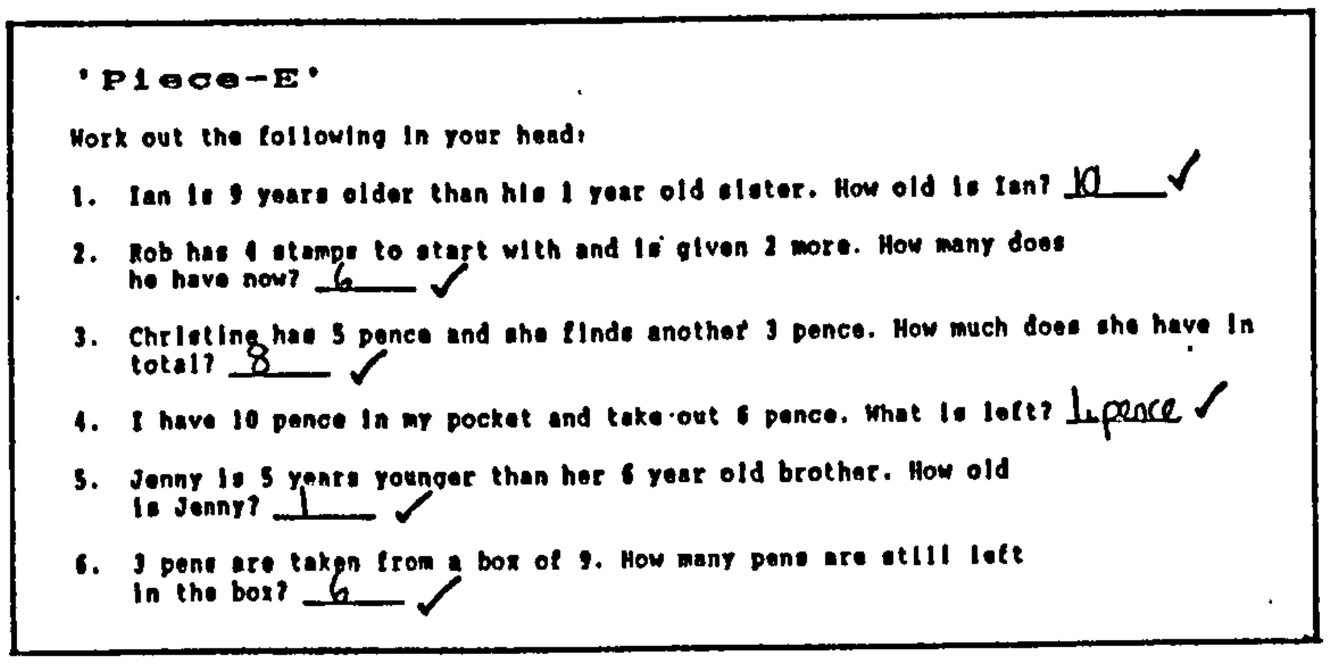
There are two key areas of interest which may be seen to emerge from the teachers' views expressed through the combined questionnaire responses. The first is the perspective teachers place upon the interaction of decision strategies (judgement policy) applied to Statements of Attainment (judgement criterion). The teachers' responses indicate Statements of Attainment to be the principal cause of concern within Teacher Assessment. However, teacher pre-conceptions, which the literature cites as influencing teacher judgements, are not acknowledged as a cause of concern. The second is the teachers' appreciation of the 'temporal' nature of sustainability. Although the views expressed by teachers did not directly mention the concept of short-term-retention; this aspect, cited within the literature more generally, is of some importance. The mismatch between teachers perceptions of sustainability and short-term-retention became of particular interest.

From the two key areas of interest it was possible to develop three aims intended for adoption within the context of this study. The first aim of this research was to investigate the interaction of teacher rating-policies applied to rating-criteria. The second aim considered the possibility of

modifying judgement policies through the use of In-Service Education and Training. The third, and final aim focussed upon the stability of judgement policies within a variety of educational contexts. These aims are now further detailed and explained.

The first aim focussed specifically on the concept of a professional judgement. Within the National Curriculum, a Teacher Assessment may be considered as the judgement policy of a teacher, applied to a student's task or activity outcome, which is referenced to one or more Statements of Attainment. Popham (1978) sub-divided the judgement process into two-stages. The initial stage is the determination of congruence. A task or activity is considered congruent with the designated assessment criterion if it 'theoretically' provides an opportunity to demonstrate attainment of that criterion. This attribute may be thought of as a pre-cursor to the second stage of the assessment process; outcome **proficiency**. The fraction or proportion of a task or activity which is required to be deemed 'correct' for attainment to be accredited to the student, is termed the proficiency level.

STUDENT'S WORK



'Piece-E'

Work out the following in your head:

1. Ian is 9 years older than his 1 year old sister. How old is Ian? 10 ✓
2. Rob has 4 stamps to start with and is given 2 more. How many does he have now? 6 ✓
3. Christine has 5 pence and she finds another 3 pence. How much does she have in total? 8 ✓
4. I have 10 pence in my pocket and take out 6 pence. What is left? 4 pence ✓
5. Jenny is 5 years younger than her 6 year old brother. How old is Jenny? 1 ✓
6. 3 pens are taken from a box of 9. How many pens are still left in the box? 6 ✓

TASK DESIGNATIONS:

proficiency status - (+)
(all answers shown are marked correct)

congruence status - (-)
(addition and subtraction facts shown up to 10 only and not 20)

JUDGEMENT SCENARIOS:

incorrect rating - (+)
(possible rating based on proficiency cue status and not congruence)

correct rating - (-)
(based on both congruence and proficiency cue status)

Figure 5.1. Student's work assessed against the judgement criterion: "know and use addition and subtraction facts up to 20".

It is evident that congruence is a necessary but not sufficient condition for the determination of a student's fulfilment of a Statement of Attainment. The issue of congruence though, is neither a commonplace term nor a well considered concept within traditional examination and assessment practices. Teachers are probably more familiar with the concept of a proficiency level; referred to as cumulative-scores or cut-scores within the literature (for e.g. Mobely, 1986). It is therefore conceivable

that teachers' judgements may covary unduly with the degree of proficiency depicted within a student's work under assessment. In these circumstances the issue of congruence would, in essence, remain redundant during the judgement process - albeit inadvertently. This is illustrated within Figure 5.1 . Although the proficiency cue is at a maximum level (ie. all correct), the task is not congruent with the Statement of Attainment. Hence, the assignment of a negative (zero) rating would be appropriate.

Within the literature the concept of assigning an inferential weighting to information in relation to its perceived salience is well documented (Shalverson and Stern, 1981). The use of such 'cognitive-simplification strategies' (Slovic and Lichtenstein, 1971) can make judgements susceptible to systematic errors. Hence, the investigation of a possible proficiency based cognitive simplification strategy employed by teachers when assessing students' work became the first aim of this study.

Aim 1. **To investigate the concept of teachers' professional judgements by considering the effects of cognitive simplification strategies on the rating process.**

The second aim addressed the broader issue of teacher professionalism and its relationship with In-Service Education

and Training. In deliberations over the nature of Teacher Assessment within the National Curriculum, acknowledgement has been made of the increased professional demands imposed upon teachers (Chard, 1990). Within the literature assessment implications focus attention specifically upon the need for the provision of effective In-Service Education and Training (Thomas, 1989). It is possible to explore the effect of such training provision on changes it may induce in teacher rating policies. This feature provided the main focus for the second aim within this study.

Aim 2. **To explore the effect on 'teacher professional judgement' of modifications to rating policies brought about through In-Service Education and Training.**

The third and final aim centred upon the issues surrounding sustainability of student performance in the context of school-based or classroom assessment. From the review of literature (GAIM, 1988) it is apparent that this concept is possibly attributable to short-term-retention effects. In order to explore this concept further it was useful to consider the alternate assessment environments available within most secondary schools. In particular, the contrast between the end-of-term test environment with that of the informal assessment conducted during the course of a lesson was considered. The

associations with short-term-retention are of differing magnitudes in both scenarios. Teachers may perceive end-of-term testing with long-term-retention, for instance. Whereas, classroom assessment may be associated with short-term-retention. The exploration of the possible effects of different classroom assessment environments on the temporal expectations of teachers and its implications for the judgement process of student performance became the concluding aim of this study.

Aim 3. **To explore teachers' perceptions of 'sustainable' student assessment performance and its relationship with short-term-retention within the context of alternate educational assessment environments.**

Leading on from the expression of the aims six hypotheses were postulated. The first three hypotheses focused upon the mechanism underlying the process of 'teacher professional judgement' and were intended to fulfil the first aim. The fourth hypothesis addressed the issue of INSET and its influence on the professional judgments of teachers and was used to fulfil the second aim. The remaining two hypotheses, although intended for the fulfilment of the third aim remained untested. The limitation of both sample size and the experimental design associated with this aspect of the research

were the deciding factors for the non-pursuance of the third aim.

The first hypothesis considered the operational characteristics of the rating policies employed by teachers during the judgement process. Congruence and proficiency were the two specific dimensions thought to constitute student cues, or attributes, within a student's work to be rated. It was assumed that cognitive simplification strategies formed the basis on which teachers professional judgements were undertaken.

Hypothesis 1. **Teachers' professional judgements are schema-based relying on cognitive simplification strategies which involve systematic rating policy errors.**

The second hypothesis centred upon the notions of information selection and interpretation. It was assumed that teachers selectively perceived and interpreted specific portions of the available information during the process of a professional judgement. Additionally, it was presumed that teachers did not recognize their utilization of this selection and interpretation process; in other words it was a heuristic strategy.

Hypothesis 2. **Teachers employ heuristic strategies in the selection and interpretation of student cue information during the rating process.**

The third hypothesis concerned the nature of the variation in rating policy adopted across teachers. The notion of an average rating policy, however, was conceptualised within the context of the potential utilization by teachers of both homogeneous and heterogeneous decision strategies. It was anticipated, though, that teachers decision strategies would be found to exist in distinct clusters.

Hypothesis 3. **Teachers' professional judgements are based on rating policies which are homogeneous.**

The remaining fourth hypothesis addressed the final aim of this study. The issue of In-Service Education and Training and its potential to modify the professional judgements of teachers was confronted within this hypothesis. The formulated hypothesis is as follows:

Hypothesis 4. **The professional judgements of teachers are significantly influenced by In-Service Education and Training.**

5.3.1 Summary.

The preliminary interview schedule and questionnaire survey results when viewed in the context of the review of literature highlight two specific and distinct themes. The first concerned the potential for the rating process to be 'schema-based' and therefore prone to systematic judgement errors. This judgement process was theorized to be susceptible to a probable heuristic strategy of selection and interpretation. The second issue considered the potential effects of In-Service Education and Training on the rating policies adopted by teachers. The exploration of such influence in the modification of rating policies was of a particular interest. Although both issues are separable through aims and formulated hypotheses, they may still be characterized as features which influence the rating policies of teachers. The next chapter will detail the empirical work undertaken within this study.

Chapter 6.

The Pilot and Main Studies:

testing the hypotheses.

The aims of this chapter are to describe the pilot and main studies undertaken within this research. This will be accomplished by describing the samples involved, outlining the measures considered suitable for testing and detailing the procedures adopted for the administration of each study. The statistical methods intended for data analysis purposes will be highlighted and their relevance to the formulated aims and hypotheses illustrated.

6.1 Introduction.

The empirical work (undertaken within this study) was intentionally of an investigatory and exploratory nature. Gross rather than fine feature effects of rating policy differences and modifications were the main focus of attention. As Popham (1981) pointed out, practical considerations of the classroom environment mitigates against the utilization of all but the most influential findings from any form of data analysis. In this context, the research design, employed within the empirical work, was designed to fulfil two specific purposes. Firstly, it had to allow the detection of teachers' baseline rating policies. The term baseline refers to the status of a rating-policy prior to a designated treatment. Secondly, it had

to incorporate and deliver several distinct treatments with the subsequent detection of any ensuing rating policy modifications.

From the research literature it was possible to isolate an appropriate research method on which to base the development of a test-instrument. In the literature 'policy-capturing' is cited as a means of determining the decision strategies employed by teachers during the process of making professional judgements (for e.g. Slovic and Lichtenstein 1971). According to this method the teacher judgement is arrived at after the integration of the available information through a process of elementary arithmetic operations (Borko and Cadwell, 1982). The resulting linear model is said to 'capture' the teachers decision making strategy or 'policy' when it can accurately predict the individuals' rating judgements.

One strategy for the construction of linear models involves the presentation of a range of hypothetical student profiles on which teachers should undertake judgements (Shalverson and Stern, 1981). These profiles may comprise of groups of students' work portfolios, incorporating systematically varied informational cues (or variables). The systematic variation of informational cues should be reflected, to some degree, within the teachers' judgements. Therefore, regression of teachers' decisions onto the informational cues could provide a basis of

a policy-capturing model. As the primary and secondary aims of this study concerned the nature and modification of teacher judgements, the utilization of a policy-capturing research technique appeared to be appropriate.

The pilot and main empirical studies undertaken within this research followed a multi-faceted data capture design and subsequent analysis. The research design was based upon a variation of a standard policy-capturing design (utilized by Borko and Cadwell, 1982). The pilot and main study work is documented within four distinct sections. The population samples surveyed; the measures employed; the administrative procedure adopted; and the data analysis techniques utilized.

6.2 The Population Samples: North Yorkshire and Humberside.

Two populations were identified from which it was thought suitable groups for the pilot and main studies could be drawn. North Yorkshire LEA provided the pilot study population and Humberside LEA the main study population. Each group completed a course questionnaire. The North Yorkshire region presented sample groups through the provision of two Mathematics National Curriculum INSET courses. The Humberside region was sampled through a process of random stratified selection and a postal questionnaire. The details of the two regional populations and their associated samples are described as follows.

Each secondary school within the North Yorkshire authority was represented at one of the two INSET courses (Appendix 2). The teacher groups available may therefore be considered as convenience samples. Additionally, the overall sample representativeness of the target population is limited by the fact the two courses provided 'intact' groups. However, it is possible to delineate certain characteristics about the two samples involved. Previous courses whose focus had been the Mathematics National Curriculum had targeted Faculty Heads and their Seconds in command. Hence, the courses used within the study had a designated population which may be characterized as those teachers whose attendance was motivated by 'interest' rather than middle-managerial obligation.

The two courses, at Grantley Hall and the York Teachers Centre, involved 30 and 34 teachers respectively. At each venue the samples were sub-divided. The allocation to these sub-groups was by random selection. Data from the Grantley Hall group was used to determine the reliability of the test-instrument reliability. Data from the York Teachers Centre group was used to test for the effect of INSET on teachers' assessment practices. Previous policy-capturing research have involved sample sizes within the range of 12 to 18 subjects. Hence, within the context of this study sub-groups of 15 and 17 subjects, were considered to be of a favourable size.

The Humberside LEA secondary schools became the population targeted for the postal questionnaire. The representativeness of the target population is enhanced by the adoption of a stratified random technique for sample selection. This technique provides more effective samples with respect to the convenience sampling approach of the North Yorkshire groups. The subject characteristic limitations attributable to the North Yorkshire groups are less of a problem with the adoption of random stratified sampling. However, the targeting of teachers within the same school produces its own 'intact' group limitations.

The stratification adopted categorized schools by taught age-range (ie. 11-16 or 11-18), and geographic area (ie. urban, sub-urban/rural). The population was sub-divided into four distinct sub-groups, one for each of the designated treatments utilized within the study (Appendix 3). A total of 32 mathematics departments were targeted; involving eight schools per sub-group (Appendix 4). With an anticipated return rate of 40%, it was expected that the requirement of 15 subjects per sub-sample (or 60 respondents in total) would be achieved. During a two month period between December 1991 and January 1992 replies were received from 48 of the estimated 160 potential respondents, a return rate of 30%. This provided an average sub-sample size of 12 subjects. This was smaller than

expected but still within acceptable limits when compared with previous research. Although an initial follow-up procedure was utilized, resource implications and time-limitations precluded the possibility of increasing the return rate with further follow-up initiatives.

In common with the preliminary investigation survey, the North Yorkshire and Humberside questionnaires included a biographic information section. The pilot questionnaire teacher sample biographic data is depicted within Tables 6.1.

Table 6.1. Teachers' biographic information (by sub-group) for North Yorkshire (sub-groups 1 to 4) and Humberside (sub-groups 5 to 8).
N=112

Service Information		<--------Sub-groups---------> 1	2*	3	4*	5	6	7	8	ST
Length of Service	0 - 4 yrs	3	3	4	4	1	1	0	2	18
	5 - 9 yrs	1	1	3	2	0	2	3	0	12
	10 -14 yrs	2	0	1	2	4	1	5	3	18
	15 -19 yrs	2	3	5	4	3	5	3	3	28
	20+ yrs	7	7	4	4	3	4	2	3	34
Principal Subject	Maths	15	11	15	11	11	12	10	11	96
	Non-Maths	0	3	2	6	0	1	3	0	15

Assessment Experience		**<--------Sub-groups---------> 1**	**2***	**3**	**4***	**5**	**6**	**7**	**8**	**ST**
Criterion - Referencing	Yes	11	9	8	2	8	12	8	7	65
	No	4	5	9	15	3	1	5	4	46
Teacher Assessment	3 yrs	7	5	9	3	6	10	9	0	49
	2 yrs	1	2	5	6	3	0	2	8	27
	1 yrs	1	1	1	3	0	1	1	0	8
	0 yrs	6	6	2	5	2	2	1	3	27

(NB. 'ST' is Sum-Total & '*' indicates missing data.)

The information documented within the biographic section of the questionnaire, depicted within Table 6.1 illustrate several features of interest across the total combination of all eight sub-groups. Possibly most noticeable is that only 16% of those sampled overall had less than 5 years teaching service; whilst 30% had in excess of 20 years. Similarly, it is of some importance to note that only 13% of respondents did not consider mathematics to be their principal subject. However, these teachers were sufficiently motivated to either attended the North Yorkshire courses or reply to the postal questionnaire. In terms of assessment experience, almost 60% of teachers were familiar with criterion-referencing techniques beyond those associated with the Mathematics National Curriculum. Of the total respondents 76% had some experience of Teacher Assessment; two-thirds of these since the introduction of the National Curriculum in 1989.

The limited size of the individual sub-groups precludes any detailed analysis or comparisons. However, there appears to be a degree of similarity between the individual sub-groups. Within each sub-group there is wide variation in length of service and assessment experience, with comparable means. There is no reason to suspect that the groups differ in any other related characteristics. Generally, there is a reasonable degree of comparability between the overall North Yorkshire and Humberside group data distributions. This, despite the use of

differing sampling techniques (ie. convenience and stratified random sampling) for the two regions.

The combined total of sub-group respondents illustrated, appear to represent a group of teachers with a considerable experience of teaching, criterion-referencing and teacher Assessment. The individual sub-sample data reflect generally the findings of the combined total of sub-groups. More specifically, a reasonable degree of similarity exists between the individual subgroups, each incorporating subjects with a variety of teaching and assessment experience. Hence, although limited in size, these subgroups would appear to provide suitable cohorts, representing the potential diversity of informed professional judgement, on which to base an investigation into Teacher Assessment.

6.3 The Experimental Measures: a description of the variables.

The empirical work undertaken within this study involved several variables which could be categorized within three distinct groups. Independent (or treatment) variables forms the first category. The second contains the control (or biographic) variables. The final category comprises the judgement (or criterion) variables. Each will now be considered and detailed further:

6.3.1 Independent Variables.

Three independent variable were involved within this study. These were associated with: a series of 'hypothetical student profiles'; an 'In-Service Education and Training package' and the 'classroom assessment environment'.

Hypothetical Student Profiles. These were constructed by the author to simulate the responses of students (age 11 to 14 years) within task activities undertaken during a Key Stage 3 Mathematics course. Three informational cues were varied: congruence of the activity with the designated assessment criteria; general proficiency of the student's response to the task; and the specific proficiency of the student's response to the task. All three cues are of a dichotomous format. The two congruence levels (yes/no) illustrate the suitability of the activity or task for assessment purposes. The two general threshold proficiency levels (yes/no) indicate the acceptability of the student's response. Finally, the two specific maximum proficiency levels (yes/no) highlight one of the extremes of a student's potential response (ie. all correct or not).

These cues were selected for two reasons. First, they represent the type of informational cues on which teachers should base their professional judgments. Second, congruence and

proficiency are mutually independent aspects of the assessment process (Popham, 1978). Hence, if teachers' professional judgements tend to covary significantly with response proficiency, rather than task congruence, this schema-based artifact should be detectable.

The individual hypothetical profiles were constructed using the following multi-stage procedure. First, since each cue had two levels, 2^3 or 8 distinct profiles could be created by the formation of all possible combinations of the three informational variables. As the general and specific proficiency cues are not entirely independent, it was possible to reduce the number of meaningful combinations to five out of eight (for eg, maximum (yes) and threshold (no) proficiency are incompatible). The utilization of five profiles ensured a comprehensive and representative subset of all possible combinations, essential in this type of research design (Edwards, 1960). Additionally, the reduction from eight to a maximum of five profiles enhanced the potential for hierarchical analysis (Airasian, Madaus and Woods, 1975), for instance Guttman Scalogram.

Second, it was necessary to select a series of Statements of Attainment on which to judge the potential hypothetical profiles. All the Statements of Attainment (Appendix 5) and their corresponding exemplars within the Mathematics National

Curriculum were studied. Multiple-criteria Statements of Attainment and those associated with exemplars of questionable utility, were eliminated to avoid confounding and ambiguity respectively. From the resulting pool of 'least ambiguous' Statements of Attainment, two batches of five criteria were identified. Each batch contained Statements of Attainment selected to differ on the key attributes of: Attainment Target; level; and requisite knowledge, skill or understanding expected of the student. This selection procedure was intended to provide a degree of independence between each of the selected criteria.

In order to select a set of congruent tasks a 'test-specification' was produced for each of the ten Statements of Attainments. Essentially, test-specifications provide the 'blue-prints' for construction (or selection) of tasks which ensure congruence with either a designated assessment criterion or a series of criteria. The production of test-specifications, within this study, followed one of the established formats utilized within the behavioural objectives movement (Popham, 1978). Subsequent inspection of contemporary secondary school texts and their associated assessment materials provided a substantial resource collection of potentially eligible tasks. Comparisons between resources (Appendix 6) and test specifications enabled the identification of tasks which were considered congruent to the chosen Statements of Attainment.

Third and finally, the series of ten selected tasks, Statements of Attainment, and their corresponding test-specifications (Appendix 7, 8 & 9) were examined by a panel of three validation judges (teachers and lecturers who had experience in educational research and the implementation of the Mathematics National Curriculum). This comprehensive examination involved two specific undertakings. The first sought the verification of task-criteria congruence. The test-specifications were included to aid this process and were themselves open to scrutiny by the judges. In particular, task-criterion congruence was determined by the judges in the context of the test-specifications and not simply the Statements of Attainment in isolation. This provided for an informed judgement with less variability and therefore greater reliability. The second required the determination of an appropriate threshold response or proficiency level for each of the tasks. The detailed results of the validation judges' deliberations indicated a general acceptance of congruence for each of the ten task-criteria pairings. Although not unanimous in their proficiency level designations, the validation judges' extensive recommendations (Appendix 10) provided sufficient information for thresholds to be determined. This was accomplished by a direct comparison of each individual judgement threshold with a pre-determined minimum *two-thirds* criterion (adopted from the CSMS research). This enabled a

series of specific thresholds to be produced which were consistent across all (acceptable) judgements.

Ten hypothetical student profiles were then constructed by systematically modifying the congruence and varying the proficiency attributes of each task and its fictitious student's response. For example, specific questions or features from a task were removed to reduce that task to a non-congruent status. Similarly, questions within a task were intentionally answered correctly or incorrectly as appropriate to illustrate a threshold or maximum level of proficiency. The two batches of five tasks were systematically modified and varied in an equivalent manner. Hence, in terms of congruence and proficiency attributes the tasks and responses formed two parallel batches. The ten tasks were allocated (arbitrarily) the alphabetic identity codings of E, S, J, Q, D (designated batch or set A) and H, A, T, O, K (designated batch or set B). The ten tasks are depicted (in full) within Appendix 14.

In-Service Education and Training. This was in the form of a training course seminar (designed and delivered by the researcher), provided at the two North Yorkshire venues and a training package incorporated within the Humberside postal questionnaire. The aim with both the training seminar and package was to address, through discussion and documentation respectively, key Teacher Assessment issues. Specifically,

these issues focussed upon the concepts of congruence and proficiency.

The concept of congruence was detailed through the identification of the task-criterion relationship. Aspects covered within this included the perceived degree of difficulty attributable to a task and the notion of its generalizability. In addition, National Curriculum exemplars were given as an aid to the interpretation of ambiguous Statements of Attainment. Finally, it was asserted that task-criterion non-congruence was more readily determined than congruence. This conjecture was justified by its analogy with the concept of proof; that is, it is 'easier' to disprove than prove. Hence, for practical purposes it was advocated that any doubt over task-criterion status should be credited as congruent.

The setting of the proficiency level was investigated through the documentary evidence of current and past assessment practices. This included information regarding the pass-marks utilized by the Examining Groups; the mark-schemes of the Key Stage 1 Standard Assessment Tasks; the criterion levels adopted within Mastery Learning; and the attainment criterion associated with the Chelsea Diagnostic Tests. There exist individual differences of policy within each of these assessment practices. However, there appears to be an aggregate proficiency level discernible from the narrow range of pass-

mark thresholds. This level, when expressed as a fraction, is 'two-thirds' and was advocated as a 'benchmark' for use by teachers in their deliberations over proficiency score thresholds.

Classroom Assessment Environments. Descriptions of classroom assessment environments were constructed to embody the concept of short-term-retention within the third independent variable. The notion of short-term-retention was conceptualised through the use of contextual information supplied with the hypothetical student profiles. Attributed to each batch of hypothetical profiles was contextual information relating to the classroom environment within which the student assessments had been 'theoretically' undertaken.

Three distinct classroom assessment environments were defined. The first was the 'everyday' classroom situation. This is probably the most popular of the three assessment environment. It provides minimal, if any, compensation for short-term-retention. The second is the classroom assessment conducted post-two-weeks of any related teaching. This strategy affords a degree of compensation; acceptable to the Graded Assessment movement for instance. The final classroom assessment environment may be conceptualised as the end-of-term test. This probably provides the most familiar and acceptable form of ***total*** compensation for short-term-retention utilized within

most secondary schools. For example, it would most likely satisfy the requirement of the Concepts of Secondary Mathematics and Science group of an unspecified but substantial time delay between teaching and testing.

The In-service and classroom assessment environment variables were represented singly or in combination through five distinct treatment conditions. These five conditions together with two additional pseudo treatments are described as follows:

Pseudo Treatments:

X_n - represented the non treatment condition, within all pretest situation;

X_o - represented the 'learning-effect' possible between pre and posttest situations;

Designated Treatments:

X_1 - represented the treatment condition associated with the use of the INSET package;

X_2 - represented the treatment condition associated with the effects of designating students' work as completed within an everyday classroom environment;

X_3 - represented the treatment condition associated with the effects of designating students' work as completed after a two-week period subsequent to related teaching;

X_4 - represented the combined treatment conditions of the INSET package with the assessment environment associated with work completed two-weeks post teaching;

X_5 - represented the combined treatment conditions of the INSET package with the assessment environment associated with work completed within an everyday classroom environment.

6.3.2 Control Variables.

Biographic Information. Detailed teacher biographic information was collected within the pilot and main study questionnaires. This data allowed the representativeness of each sub-sample to be considered. It was then possible to judge each samples suitability for use within the study (see p106 for sub-sample analysis). More specifically, this section of the questionnaire gathered information about teaching service, subject specialism and both previous and current criterion-referenced experience. It was anticipated that the function of teachers' professional judgments would covary with one or more aspects of the background information. In this respect, the teacher characteristics considered were thought eligible for use as control or moderator variables.

6.3.3 Judgement Variables.

Criterion-Task Judgements. Subjects were asked to judge each hypothetical profile against its corresponding Statement of Attainment. The judgement process involved the determination of criterion fulfilment for each hypothetical profile and its corresponding Statement of Attainment. This was recorded dichotomously; the profile was deemed either to fulfil the Statement of Attainment or to be deficient. This dichotomous format was adopted throughout the pilot and main study. However, two sub-groups within the main study utilized an enhanced format, although still dichotomously recorded. The determination of criterion fulfilment was sub-divided into two distinct components; those of congruence and proficiency. On these occasions, each subject had to determine the level of fulfilment for each criterion separately.

Assessment Profile Ratings. Subjects were asked to provide a series of judgements about several criterion-referenced assessment aspects within the pilot study questionnaire. These judgements were intended to provide explicit parameters within which implicit decision strategies, applied to hypothetical profiles, could be studied. The five specific assessment aspects considered were related to:

(1) The expectation that a proficiency difference between informal classroom and formal test based assessment should exist (yes or no);

(2) The anticipated threshold value (expressed as a fraction or proportion) at which the proficiency difference between informal and formal classroom based assessment conditions should occur (eg. 4/5 = informal classroom assessment and 3/5 = formal classroom assessment);

(3) The determination of a proficiency rating, expressed as a quartile value, for both informal and formal classroom based assessment (eg. 4/4 = informal classroom assessment 3/4 = formal classroom assessment);

(4) The estimation of a minimum period of time after which short- term-retention should have no significant or measurable effect (eg. 1 week, 2 weeks);

(5) The acceptance that the majority of SoAs (designated as a minimum of 75%) were adequate for the purpose of Teacher Assessment (yes or no).

6.3.4 Eliminated Variables.

Although all of the variables involved within this study have been described during the previous section, it is important to list a collection of variables of value if only due to their absence. These eliminated variables will now be detailed:

Gender. Within the literature gender is believed to have an insignificant biasing effect on the rating process (Dusek and Joseph, 1983). However, to eliminate any possibility of confounding effects with the numerous other variables involved within the study, this variable was made redundant from the hypothetical profiles by the intentional absence of identifying student names.

Prior Assessment Influences. Previous studies have indicated the possibility that teachers may attain a consistent description of the student by allowing ratings associated with earlier assessed work to influence successive judgements (Archer and McCarthy, 1988). Compensation for this effect was achieved by creating an artificial independence across the profiles, ie. each of the ten profile constituted the work of as many individuals.

Attainment Expectation Effects. National Curriculum Level information and an indication of an individual student's set or group designation, if augmented to the student profiles, were thought to present a possible biasing influence on the judgement process (Hoge and Butcher, 1984). The hypothetical profiles were intended to be assessed as presented and without the influence of any artificial expectation generated as a consequence of any attainment level or set/group labelling.

Hence, exclusion of such labelling from the student profiles minimized the consequences of any associated problems.

Ambiguous Criteria Effects. Within the National Curriculum literature the problems associated with ambiguous Statements of Attainment are detailed at length (for eg. SMP, 1990). The intentional avoidance of multiple-criteria Statements of Attainment eliminated one aspect of this ambiguity. The test specifications constructed for those criteria allied to the hypothetical profiles enabled any further ambiguities to be identified and eliminated from the selection process.

6.3.5 Test Instrument Format.

The basic test instrument had four distinct sections. The first, in the form of a questionnaire (Appendix 11), requested specific biographic details including number of years teaching and relevant experience of Teacher Assessment. The second component, again in the form of a questionnaire (Appendix 11), focussed upon issues relating to Teacher Assessment. Three assessment characteristics were examined. These were: the adequacy of Statements of Attainment within the determination of task-criterion congruence; the effect on proficiency ratings of assessment context (ie. classroom environment); and the temporal nature of short-term-retention. The third component involved the presentation of two sets of hypothetical profiles.

These were depicted as reduced photocopies of students' work each illustrating a specific mathematical task and its appropriately fabricated solution. The fourth component listed the Statements of Attainment, each attending to a specific hypothetical profile, together with information regarding the completion of the judgement process. Additional, brief contextual information indicated the circumstances within which the tasks (illustrated within the student profiles) were undertaken. The professional judgement of criterion with hypothetical profile was recorded as a dichotomous (yes/no) response within this latter component. The fifth and final component was of an instructional format; documenting the In-Service Education and Training package. This detailed information relating to both issues of congruence and proficiency and was intended to provide guidance for teachers with respect to the judgement process. From the basic test instrument three versions were developed:

Version 1. This version was utilized within the initial phase of the pilot study and included an A4 'flysheet' and A3 'foldover' document. The A4 'flysheet' contained the two components associated with teacher biographic information, and Teacher Assessment issues. The A3 'foldover' document comprised two components; namely the hypothetical profiles and the Statements of Attainments. Each side of the document detailed one set of five profiles and their allied judgement criteria.

Additionally, this version existed in two parallel forms; each presenting the two sets of profiles in a parallel but reverse order.

Within this version all ten hypothetical profiles were utilized twice, in consecutive formats. These were test (Appendix 12) and re-test (Appendix 13) configurations, the latter allowing the determination of test instrument reliability. The presentation of Statements of Attainment, and therefore hypothetical profiles, were re-ordered in a pseudo-random sequence on the post-test format to that of the pre-test.

Version 2. This second version was adopted within the second and final phase of the pilot study and also included an A4 'flysheet' and A3 'foldover' document. In common with the first version, the A4 'flysheet' contained the two components associated with teacher biographic information, and Teacher Assessment issues. Similarly, the A3 'foldover' document comprised two components; namely the hypothetical profiles and the Statements of Attainments. Similarly, parallel forms were created which presented the two sets of five profiles in reverse order.

Within this version however, the ten hypothetical profiles were utilized singly in sets of five. These were pre- and post-test configurations, the latter aimed at determining the

effectiveness of the trial In-SET package utilized. Hence, in this version only one side of the A3 document was detailed with one set of five hypothetical profiles and their allied Statements of Attainment.

Version 3. This third and final version was utilized within the main (postal) study and, as with previous versions, adopted an A4 'flysheet' and A3 foldover document. The pilot study undertaken to establish the reliability of the test-instrument and the effectiveness of the In-Service Education and Training package indicated these to be both acceptable and appropriate for use within the main study (chapter 7 will detail the reliability measures and the INSET treatment analysis further). In deference, therefore, to the previous versions, the constituent components of the test instrument remained unchanged; although their configuration was altered. The A4 'flysheet' (Appendix 14) contained the two sets of hypothetical profiles (one set of five on each side). In this version the two sets of five profiles were presented in a unique order; no complementary format was required. The A3 'foldover' document comprised three components on this occasion (Appendix 15). The first requested teacher biographic information through a questionnaire format. The second presented In-Service Education and Training materials. The third component detailed the two sets of five Statements of Attainment (associated with hypothetical profiles).

In common with the second version, the ten hypothetical profiles were utilized singly in sets of five within consecutive formats. These were pre- and post-test configurations, again the latter aimed at determining the effectiveness of the In-Service Education and Training package utilized. Post-test security was achieved by the use of a paper-clip (attached to the A3 'foldover' sheet) to enclose both the In-Service Education and Training materials and the commensurate set of Statements of Attainment (Appendix 16). Finally, within this version the instructional information associated with the set of five post-test Statements of Attainment presented four distinct alternative assessment contexts. Hence this latter version was essentially four distinct sub-versions.

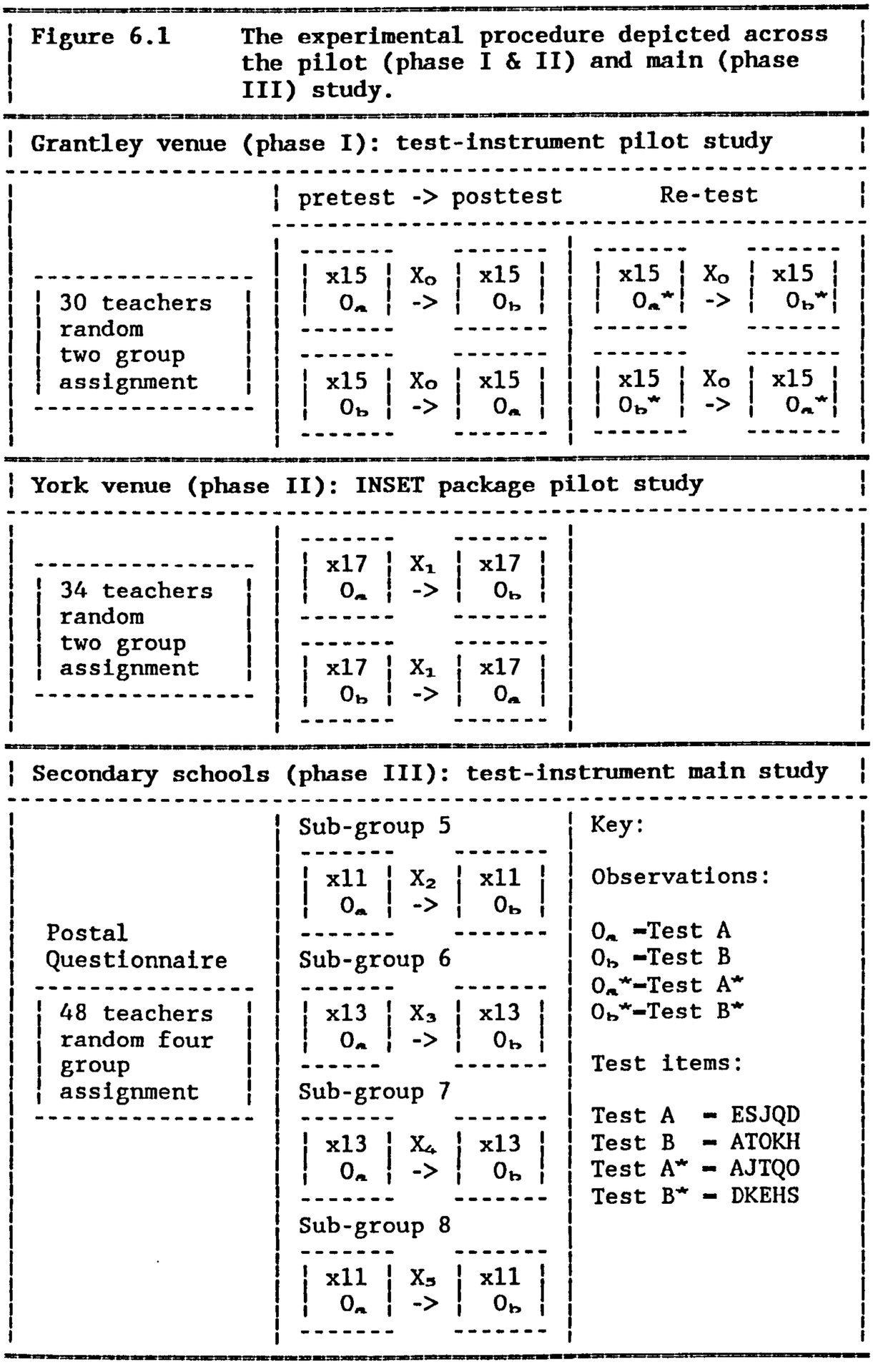
The test instrument described appears to represent a range of independent, control and judgement variables. Additionally, despite the utilization of only a limited number of hypothetical profiles within the study the main underlying dimensions of congruence and proficiency are explored in detail. The pilot study established the reliability of the test-instrument and the effectiveness of the trial In-Service Education and Training materials. This allowed, therefore, their adoption across the main phase of the study; with only configurational changes. The co-option of 'judges' allied to

the use of test specifications during the construction of the hypothetical profiles ensures validation of the overall process and outcome. Although only three validation judges were involved, the extensive nature of their brief enabled a test instrument to be constructed which was both comprehensive and appropriate for the purpose to which it was intended.

6.4 The Experimental Procedure: the pilot and main studies.

The experimental procedure was undertaken in three phases this is illustrated within Figure 6.1 . The first two phases involved the pilot study. The first of these aimed to measure the reliability of the test-instrument utilized within the research. The aim of the second phase was to determine the mechanisms by which teachers professional judgments are undertaken. In addition, this investigation was accompanied by the use of an In-Service Education and Training package whose purpose was to seek to modify the judgement mechanisms utilized by teachers. The third and final phase was centred upon the main (postal) study and was aimed at complementing the pilot study. This involved the further exploration of teachers' professional judgements of hypothetical profiles (involving materials associated with alternate classroom assessment environments and an additional In-Service Education and Training package). Each phase will now be detailed more fully.

Figure 6.1 The experimental procedure depicted across the pilot (phase I & II) and main (phase III) study.



6.4.1 Phase I: pilot study.

The first phase of the pilot study was undertaken within Grantley Hall (North Yorkshire) during November 1991. Teachers (30 in all) were randomly assigned to one of two groups. Each subject participated individually in the study utilizing version 1 of the test-instrument. At the beginning of the session a brief introduction was given indicating that the purpose of the study was to investigate Teacher Assessment and, more specifically, the process of teachers professional judgements within the Mathematics National Curriculum. No mention of schematic and/or heuristic decision making strategies was made either implicitly nor explicitly.

The subjects involved participated in two sessions, the first at the start of the day. The second was undertaken some 6 hours later at the end of the day and constituted a re-test. On both occasions, the subjects were instructed to make *no assumptions* regarding the student profiles presented and were reminded that each of the ten profiles represented the work of as many individual students. Each group was allocated one of two parallel forms to rate (ie. each group rated the two sets of five profiles in a reverse order). Subjects were expected to complete the two sets of five ratings within a 10 minute time period (ie. 10 minutes per session - Appendix 17). Although during the end of day session the two sets of profiles were

psuedo-randomly re-ordered, this did not require alternative administrative procedures to that of the morning. The re-ordering was undertaken to minimise any reactive effects of the initial rating exercise (Campbell and Stanley,1966). By comparing the response patterns of sub-groups with complementary presentation formats it was possible to estimate the magnitude of any such effects; although compensation was not a viable option.

The completion of the end of day re-test was followed by a short plenary session which included the administration of a questionnaire. The session focussed upon the issues of congruence and proficiency; each was defined for the benefit of the participants. The questionnaire required a series of responses concerning biographic details and issues of assessment characteristic ratings (including those related to aspects of congruence and proficiency).

6.4.2 Phase II: pilot study.

The second phase of the pilot study was conducted at the York Teachers' Centre (North Yorkshire) again during November 1991. Teachers (34 in all) were randomly assigned to one of two groups. As with the first phase proceedings, each subject participated individually within the study utilizing version 2 of the test-instrument. At the beginning of the session a brief introduction was given indicating the purpose of the study and

again mention of schematic and/or heuristic decision making strategies was neither made implicitly nor explicitly.

The completion of the hypothetical profile ratings was preceded by a short administrative briefing identical to that given at the first phase. However, unlike the first phase only a single set of five hypothetical profiles was presented at each session (again separated by some 6 hours). Hence, across the two sessions all ten profiles were encountered. Each group was allocated one of two parallel forms to rate (ie. the two groups rated the sets of five profiles in a reverse order across both sessions). Again, subjects were expected to complete the two sets of five ratings within a total time period of 10 minutes (ie. five minutes per session).

The In-Service Education and Training package was delivered prior to the end-of-day testing session. The package was approximately 30 minutes in length and involved the presentation of factual information relating to congruence and proficiency. This was delivered via an Over-Head Projector transparency and a follow-up discussion (Appendix 18). After the end-of-day test an identical questionnaire to that administered during the first phase was completed by the participants. The In-Service Education and Training package was based upon the debrief materials utilized within the first

phase. Hence, a debrief was not a pre-requisite to the completion of the questionnaire in this second phase.

6.4.3 Phase III: main study.

The third and final phase which formed the main study was undertaken within a stratified random sample of 32 Humberside secondary schools. A collection of several test-instruments and accompanying administrative instructions, addressed for the attention of the Head of Mathematics, were despatched to each school within the sample. Due to the postal nature of the main study all preliminary information relating to the purpose of the work and the necessary administrative procedures had to be in a documentary form. These were introduced within a covering letter initially requesting the co-operation of the Head of Mathematics and their associated teaching staff. The administrative procedure information conveyed was almost identical to that given verbally during the first two phases. As with the earlier phases, each subject was expected to participate individually within the study (which utilized version 3 of the test-instrument). As in phases I and II no mention of schematic and/or heuristic decision making strategies was made either implicitly or explicitly within the despatched materials.

Completion of the hypothetical profiles followed the format adopted within phase two. Each set of five profiles was

undertaken in one of two sessions (although no prescribed separation period was advocated). On this occasion, the presentation order of profile sets was identical for all participants. Four sub-versions of the test-instrument were available for use within the main study. However, within each school cohort only one sub-version was utilized. In each case explicit instructions accompanied the rating materials to enable the completion of each aspect of the test-instrument in the correct sequence. The two sets of rating materials occupied alternate sides of the A3 'foldover' document and a paper-clip ensured the correct order of completion was undertaken. Additionally, this measure ensured security for the In-Service Education and Training package printed within the A3 'foldover' document itself.

The procedures adopted had three aspects which could potentially mitigate against the success of this study. These were the extensive nature of the administrative information utilized within the test-instrument; the complexity of the rating and questionnaire materials; and the confidentiality of the In-Service Education and Training package in the main study. The success of the procedures adopted within each phase is evident, however, from the comprehensive detail of the data collected. Of those rating forms and questionnaires completed and examined, less than 1% of the requested information was omitted across the range of 112 participating subjects.

Similarly, the separate ratings expected for congruence and proficiency within version 3 enabled the ability of the test-instrument to capture the required data to be evaluated. Of the 240 ratings (for the two sub-groups) less than 3% indicated a mismatch between the combination of congruence and proficiency ratings with the overall judgement.

6.5 The Data Analysis: an overview of statistical methods.

The analyses undertaken within this study were designed to fulfil three criteria. These were the identification, classification and evaluation of teacher decision strategies. The first two criteria were fulfilled through an analysis centred upon the exploration of teacher decision strategies. The final criteria was achieved by investigating the stability of these decision strategies within different treatment conditions and assessment contexts. The original intention was to incorporate, within the overall statistical consideration of the data, an analysis of covariance. However, the response patterns associated with the two sets of five profiles did not covary sufficiently within the first phase of the pilot study to allow the eventual adoption of this technique. Although, the early acceptance of test-instrument reliability (p146) enabled the adoption of this across the remaining phases of the pilot and main studies. Hence, the various analyses utilized could

compare data (where possible or appropriate) across all three phases of the research.

The analysis of data was undertaken on three levels. The first was of a preliminary nature; providing an ***overview of the empirical data*** including the calculation of descriptive statistics for all sub-groups. The second level focussed upon a ***within sub-groups analysis*** employing a common set of statistical techniques applied to the data. These were utilized to establish the initial relationship of the data to the aims and hypotheses. Finally, the third level adopted a common set of statistical techniques for a ***between sub-groups analysis*** (ie. compared across several sub-groups). This provided confirmatory and supportive evidence for the relationships established in the second level of data analysis. Each level of analysis will now be discussed in further detail.

Overview of the empirical data. The descriptive statistics calculated for each sub-group included response frequency totals; means and standard deviations. These were collated to enable an overview of the empirical data to be established and to allow within and between sub-group comparisons to be undertaken. It was anticipated that the descriptive statistics alone would only highlight treatment effects of a considerable magnitude. A further analysis was deemed necessary if the

detection of more detailed evidence, concerning less appreciable treatment effects, was to be accomplished.

Within sub-groups analysis. For each of the two pilot and single main study phases, the data analysis followed a fixed sequence of statistical procedures. The first procedure involved the use of regression analysis; and provided the initial exploration of the data. The remaining correlational and ordering-theoretic hierarchical analyses presented opportunities for additional, complementary evidence to be collected.

The first procedure, a linear regression analysis, was performed on the data collected for each sub-group sample. Each subject's ratings were regressed onto the three underlying dimensions associated with the hypothetical profiles. The resulting regression coefficients and equation were considered to 'capture' the subject's judgement or rating policy (Borko and Cadwell, 1982). In addition, average regression equations were calculated for collective sub-group data.

Inspection of the calculated regression coefficients was undertaken to determine the degree of between-subjects homogeneity within each sub-group sample. If this determination indicated the assumption of a common rating policy was untenable, then consideration was given to other possible

descriptions of variations among teachers' decision strategies. If the variation appeared random about an average value then it would be appropriate to adopt the overall sub-group regression equation. Alternatively, if the subjects utilized systematically different strategies then reporting an 'average' policy would be inappropriate. In this latter circumstance an attempt was made to 'cluster' teachers into groups with homogeneous decision strategies and to estimate one set of regression coefficients for each of these.

Next, two sets of correlation matrices were calculated for each of the sub-groups. The first set considered the inter-correlations of the response patterns associated with the ten hypothetical profiles for each sub-group. The second focussed upon the inter-correlations of the regression weights associated with the three underlying dimensions. Inspection of the two sets of correlation matrices were undertaken to determine the magnitude of the covariation between the hypothetical profiles and the three underlying dimensions for each sub-group. Hence, it was possible to investigate the mechanism through which the schema based judgement process could be characterized. These analyses, carried out across all sub-groups, enabled the sensitivity of the observed covariations to be investigated within the context of differing treatment conditions.

Finally, an ordering-theoretic hierarchical analysis statistical technique was utilized for the within sub-group analysis. For each sub-group a pre-requisite hierarchy was calculated from the response data. Having established a hierarchical relationship within each of the hypothetical profile sets it was possible to investigate the schema based judgement process within this context. In particular, the sensitivity of the hierarchies could be explored across the differing treatment conditions.

Between sub-groups analysis. For each of the two pilot and single main study phases, the data analysis again followed a fixed sequence of statistical procedures. Initially, this involved inspecting the collective regression data from the within sub-groups analysis in overview. This was intended to provide confirmatory evidence for the findings of the within sub-groups analysis. The remaining regression and then cluster analyses provided opportunities for evidence supportive of previous findings to be collected.

The initial procedure, involving the inspection of the regression data from the within sub-groups analysis, was intended to confirm the stability of similarity and difference patterns discovered within the individual sub-group response data. Three distinct features were sought. The first focussed upon the variation of decision strategies adopted by teachers.

Each sub-group was characterized by its range or collection of distinct decision strategies (associated with the pre-treatment condition); these distributions were then compared on an individual sub-group basis. Further, broader comparisons were then undertaken using sub-group 'average' decision strategies as the basis of the analysis. The second considered the effect of treatments on decision strategies adopted during the rating process. More specifically, comparisons were made between sub-groups in post-treatment conditions. Again sub-group decision strategy distributions and averages formed the basis of this analysis. The third centred upon the differences between the two sets of five hypothetical profiles. Comparisons were made between sub-group decision strategy distributions for the two sets of profiles.

Next, the regression coefficients obtained from the within sub-groups analysis of subject ratings and underlying dimensions (three informational cues) were themselves regressed onto the assessment profile ratings. If the resulting regression equation provided an inconsistent model for the prediction of the underlying dimensions then this could indicate the possibility of a heuristic decision strategy in operation. That is, teachers may be unaware of the differences which may exist between their perception of assessment in theory and their manifestation in actual practice. The identification of any potential moderator variables which could account for

individual subject decision strategy differences involved the use of a second regression analysis. The regression coefficients obtained from the within sub-groups analysis of subject ratings and underlying dimensions were regressed onto the teacher biographic information (identified as potential moderator variables). If the resulting regression equation provided a consistent model for the prediction of the underlying dimensions then this could indicate the utility of certain biographic information as potential moderator variables.

Finally, hierarchical clustering analyses were performed for sub-group data. The intention was that this procedures should support the previous regression, correlational and ordering-theoretic analyses. The techniques utilized the rating policy regression coefficients calculated during the within sub-groups analysis. The overall aim was to identify and classify (hierarchically cluster) each of the common decision strategies both within and across sub-groups. The procedure adopted encompassed three distinct aspects. The first focussed upon the pre-treatment sub-groups with the identification and characterization of different decision strategies (both within and between sub-groups). The second considered post-treatment decision strategies, again within and between individual sub-groups. The third aspect compared the clustering of decision strategies for the two profile sets (ie. between the two sets

of five hypothetical profiles) in the various pre- and post-treatment scenarios. It was expected that the trends and patterns identified in the previous analyses would be of a magnitude sufficient for identification and classification purposes using this statistical technique.

6.6 Internal and External Validity.

The research design utilized within this study was intentionally exploratory. It was anticipated that the major effects associated with teachers' professional judgements would be of a sufficient magnitude to enable their detection with a policy-capturing design. The sample sizes adopted were equal to or slightly smaller than those for comparable studies of this nature. However, sample size alone may not be taken as a predictor of a data set's utility; sample representativeness is of greater significance. Hence, within this study it was sought, through the use of appropriate sampling techniques, to estimate the representativeness of each sub-group. Thus, it was possible to gauge both the utility and limitations associated with each sample.

The use of validation judges ensured the policy-capturing test-instrument developed was appropriate for its purpose. The detailed delineation of each hypothetical profile, in terms of congruence and proficiency, enabled the construction of a test-

instrument capable of minimizing potential threats to external and internal validity. The threat to external validity of the reactive effect of testing was identified by adopting parallel experimental arrangements with the rating of hypothetical profiles sets presented in a complementary order. Although identification of the reactive effect of testing was possible, compensation for this was not.

Unlike the factors jeopardizing external validity, which were identified but not compensated for, the threats to internal validity were more successfully minimised for. Firstly, the test-instrument design eliminated the main sources of potential bias through its selection of measures (or variables). Both history and maturation considerations were rendered obsolete as both the pre- and post-treatment testings were undertaken in the same or consecutive sessions. In the latter case, full account was possible of the events occurring between the consecutive sessions. For each of the instances in question no significant historical effects (and therefore sources of invalidity) were identified. The effect of testing was considered to present a significant threat to the validity of the experimental procedure. In common with the reactive effects of testing, the adoption of parallel experimental arrangements, with the rating of hypothetical profiles sets presented in a complementary order, allowed the identification of testing effects. In respect of the re-test arrangements any appreciable

testing effects could invalidate the determination of test-instrument reliability. However, these measures of reliability need to be seen within the context of the full data analysis outlined in chapter 7 and discussed at length during chapter 8.

6.6.1 Summary.

The extensive nature of the measures utilized and procedures adopted within this study enabled the development of a test-instrument capable of collecting an extensive range of data. The nature of the multi-faceted data sets collected, compelled the utilization of an equally comprehensive series of statistical techniques for their analysis. Initially, these focussed upon a series of exploratory analyses; designed to test the viability of the proposed hypotheses. Finally, confirmatory analyses were undertaken, these provided supportive evidence for the viability (or not) of hypotheses explored previously. The results of these analyses, including the contextual measures surrounding the test-instrument reliability data, will be presented and discussed in the next chapter.

Chapter 7.

The Results of the Pilot and Main Studies: an analysis of the data.

This chapter describes the results of the statistical analyses applied to the data collected within the pilot and main studies. An initial overview of the data will be followed by a range of comprehensive 'diagnostic' analyses. The relationship between the results and the associated aims and hypotheses will then be identified and briefly discussed. Finally, these findings will be described and summarised to enable the aims and hypotheses to be evaluated.

7.1 Introduction.

The data collected within the pilot and main studies was incorporated within a 'policy-capturing' research model which required the use of statistical techniques appropriate for a multi-dimensional analysis. This series of diagnostic analysis enabled the nature of the decision making process to be investigated. However, it was also possible to apply a series of uni-dimensional analyses to the cumulative scores available within the data. These cumulative scores could be interpreted as a measure of criterion-referencing skill or ability. These uni-dimensional analyses were not inherent in the adopted policy-capturing research design and were therefore expected to

be of only limited value within the context. For instance, the construction of the profile-sets was not undertaken with strict consideration given to conventional test-item selection. Similarly, conventional test-item discrimination criteria were not adopted. Uni-dimensional protocols were unnecessary in the construction of profiles appropriate for a multi-dimensional policy-capturing research model.

The nature of the two sets of five profiles; each incorporating three underlying dimensions necessarily precludes any simplistic data manipulation and subsequent comparisons. Instead any score variations need to be viewed in terms of judgement policy differences. These differences are manifest in the response to individual profiles and have specific meaning which a cumulative score is unable to reflect. A score of 4/5 may be associated with any one of the following series of response patterns:

01111 10111 11011 11101 11110

Each pattern representing a combination of particular underlying dimensional influence and commensurate judgement policy. The multi-dimensional analysis, associated with the policy-capturing research model, provides a diagnostic measure of each response pattern.

Although of limited value the results of a uni-dimensional analysis were considered to be important for two reasons.

Firstly, it would provide preliminary findings and indicate treatment effects (of a significant magnitude). Secondly, any relationship between the uni-dimensional and multi-dimensional analysis findings could be examined and inconsistencies identified. The preliminary analysis includes tabulation of response score frequencies, item facility values, oneway analysis of variance and t-tests. Collectively, they provide an overview of the data.

7.2 An overview of the data.

The data were organised within two distinct formats (Table 7.1). The first depicted the *actual* responses given by the teachers within the profile section of the test instrument. This involved tabulating the *Yes/No* ratings, given to each profile, as corresponding 1/0 values (ie. Yes=1; No=0). The second indicated the response *adjusted* for their accuracy. This involved re-tabulating the *Yes/No* ratings as correct or incorrect, again using 1/0 values (ie. correct=1, incorrect=0). For example in the *actual* response format the ratings of Yes, Yes, No, No, No for profiles E, S, J, Q, D respectively would be depicted as the pattern 11000. In the *adjusted* response format this would be given as 00100. The latter indicating a cumulative score of 1 (out of 5).

Table 7.1 Response patterns in *actual* and *adjusted* formats for Profile-Set A (PSA). N=15

Subject	*Actual* Response Pattern					*Adjusted* Response Pattern					CT Score
1	1	1	1	1	1	0	0	0	1	1	2
2	1	1	0	1	1	0	0	1	1	1	3
3	0	1	0	1	1	1	0	1	1	1	4
4	0	1	0	1	1	1	0	1	1	1	4
5	1	1	0	1	1	0	0	1	1	1	3
6	1	0	0	0	1	0	1	1	0	1	3
7	1	1	0	1	1	0	0	1	1	1	3
8	1	1	0	1	1	0	0	1	1	1	3
9	1	1	0	1	1	0	0	1	1	1	3
10	1	1	0	1	1	0	0	1	1	1	3
11	0	1	0	1	1	0	0	1	1	1	3
12	0	1	1	1	1	1	0	0	1	1	3
13	1	1	0	1	1	1	0	1	1	1	4
14	1	1	0	1	1	0	0	1	1	1	3
15	1	0	0	1	1	0	1	1	1	1	4

(CT Score = Cumulative Score)

Reliability (test-retest analysis)

Within sub-groups 1 and 2 the use of the test-battery pair in a re-test situation enabled a reliability analysis to be undertaken. Phi and percentage agreement values were calculated for both sub-groups and profile-sets. The results of the analysis yielded phi coefficients of .85 ($p<.01$) and .78 ($p<.01$) for Profile-Set A (ESJQD) within sub-groups 1 and 2 respectively. Associated percentage-agreement figures were 93% and 89%. Similarly, phi coefficients of .86 ($p<.01$) in both cases were recorded for Profile-Set B (ATOKH) within sub-groups 1 and 2. Corresponding percentage-agreement figures were 93% in both cases. The level of significance of the phi coefficient values and the supportive percentage agreement scores were

considered to be an adequate indicator of the test-battery's reliability.

Table 7.2. Descriptive measures of cumulative score frequencies for PSA and PSB.

Sub-group	N=	Tr't	<----- Score Frequencies PSA -----> (0)	1	2	3	4	5	Min	Max	Range
1	15	X_n	0	0	1	10	4	0	2	4	2
2	15	X_o	0	0	1	5	8	1	2	5	3
3	17	X_n	0	0	3	3	8	3	2	5	3
4	17	X_1	0	0	0	6	6	5	3	5	2
5	11	X_n	0	0	4	3	3	0	2	4	2
6	13	X_n	0	0	1	6	5	1	2	5	3
7	13	X_n	0	0	2	3	7	1	2	5	3
8	11	X_n	0	0	0	3	5	3	3	5	2

Sub-group	N=	Tr't	<----- Score Frequencies PSB -----> (0)	1	2	3	4	5	Min	Max	Range
1	15	X_o	0	1	0	1	13	0	1	4	3
2	15	X_n	0	0	0	3	11	1	2	5	3
3	17	X_1	0	3	1	5	7	1	1	5	4
4	17	X_n	0	1	4	5	3	4	1	5	4
5	11	X_2	0	1	1	5	3	1	1	5	4
6	13	X_3	0	0	3	4	6	0	2	4	2
7	13	X_4	0	1	1	3	8	0	1	4	3
8	11	X_5	0	0	1	4	6	0	2	4	2

(Tr't = treatment condition)

For each profile-set, ***adjusted*** cumulative score frequencies were calculated (Table 7.2). Inspection of Profile-Set A (PSA) across sub-group 1 revealed an absence of scores of zero and 1, and a degree of clustering around the values of 3 and 4. There appeared to be no consistent pattern across the range of the X_o treatment groups. Profile-Set B (PSB) illustrated, again, an absence of the extreme score of zero, and a similar degree of clustering around the values of 3 and 4. Despite the absence of

low values, the distributions represent the full range of scores. Generally, the distribution of scores displayed an acceptable level of sensitivity within the context of this preliminary analysis.

Table 7.3 Descriptive measures of cumulative score frequencies for PSA and PSB.

Sub-group	N=	Tr	<----- Cumulative Scores PSA -------> Mean	StDv	Kurt	Skew	Min	Max	Range
1	15	X_n	3.20	0.56	0.3	0.1	2	4	2
2	15	X_o	3.60	0.74	0.4	-0.4	2	5	3
3	17	X_n	3.65	1.00	-0.6	-0.5	2	5	3
4	17	X_1	3.94	0.83	-1.5	0.1	3	5	2
5	11	X_n	2.91	0.83	-1.5	0.2	2	4	2
6	13	X_n	3.46	0.78	0.2	0.2	2	5	3
7	13	X_n	3.54	0.88	-0.1	-0.6	2	5	3
8	11	X_n	4.00	0.77	-1.1	0.0	3	5	2

Sub-group	N=	Tr	<----- Cumulative Scores PSB -------> Mean	StDv	Kurt	Skew	Min	Max	Range
1	15	X_o	3.73	0.80	11.4	-3.3	1	4	3
2	15	X_n	3.87	0.52	1.4	-0.3	2	5	3
3	17	X_1	3.12	1.22	-0.4	-0.7	1	5	4
4	17	X_n	3.29	1.26	-1.0	0.0	1	5	4
5	11	X_2	3.18	1.08	0.8	-0.4	1	5	4
6	13	X_3	3.23	0.83	-1.3	-0.5	2	4	2
7	13	X_4	3.38	0.96	2.1	-1.6	1	4	3
8	11	X_5	3.45	0.69	0.1	-0.9	2	4	2

Descriptive measures were calculated for the *adjusted* response format data and are illustrated within Table 7.3. These measures utilised the cumulative scores found within each sample and profile-set respectively. Again within the context of this preliminary analysis the skew and kurtois values of the cumulative score frequencies were generally considered to be of

an acceptable size. This was supportive of the earlier sensitivity findings of the *adjusted* cumulative score distribution analysis (Table 7.2).

The relationship between the responses to different profiles was examined by comparison of their individual facility values (Table 7.4). Within PSA and PSB the range of facility values was found to be extensive (0.00 to 1.00). Profile D (PSA), for instance, tended to be correctly rated by the majority of subjects, irrespective of the sub-group. Conversely, within PSB Profile S was more difficult to rate correctly, again a pattern consistent across all sub-groups. These examples indicate the inability of certain profiles to adequately discriminate between subjects of high and low ability. The range of facility values within both profile-sets indicated one limitation associated with these as uni-dimensional measures. Similarly, the facility values for profile-pairs (for example E and H), in theory matched in terms of their multi-dimensional status, did not provide a consistent pattern for equal treatment conditions. Another example of a limitation of the profile-sets as multi-dimensional measures.

Table 7.4. Descriptive item response score facilities for PSA and PSB.

Sub-group	N=	Tr	<------- Response Scores PSA -------> E	S	J	Q	D	Mean
1	15	X_n	0.27	0.13	0.87	0.93	1.00	0.64
2	15	X_o	0.73	0.20	0.73	0.93	1.00	0.72
3	17	X_n	0.59	0.41	0.82	0.82	1.00	0.73
4	17	X_1	1.00	0.53	0.82	0.71	0.88	0.79
5	11	X_n	0.45	0.00	0.45	1.00	1.00	0.58
6	13	X_n	0.92	0.23	0.62	0.69	1.00	0.69
7	13	X_n	0.77	0.15	0.77	0.92	0.92	0.71
8	11	X_n	0.91	0.27	0.82	1.00	1.00	0.80

Sub-group	N=	Tr	<------- Response Scores PSB -------> H	A	T	O	K	Mean
1	15	X_o	0.00	0.93	0.93	0.87	1.00	0.75
2	15	X_n	0.07	1.00	0.80	1.00	1.00	0.77
3	17	X_1	0.24	0.82	0.88	0.53	0.65	0.62
4	17	X_n	0.24	0.76	0.53	0.88	0.88	0.66
5	11	X_2	0.09	0.73	0.45	1.00	0.91	0.64
6	13	X_3	0.15	0.85	0.77	0.54	0.92	0.65
7	13	X_4	0.15	0.85	0.77	0.77	0.85	0.68
8	11	X_5	0.18	0.81	0.55	0.91	1.00	0.69

T-test (analysis of means)

After the inspection of *adjusted* response score distributions, descriptive measures and individual profile facility values, the analysis of the data concluded with a series of significance tests. Firstly, individual sub-group mean differences (between profile-set pairs) were examined with a series of t-tests, and are represented within Table 7.5. These were intended to provide evidence of a 'learning effect' (treatment condition X_o) present as a consequence of pre-testing. The t-test determines where the means of selected sub-group pairings differ sufficiently to cause rejection of the

..

null-hypothesis that they are members of the same population. Non-significant results were recorded for both instances of X_o associated with PSB and PSA, within sub-groups 1 and 2 respectively. These results indicated the probable absence of any 'learning effect'; the reverse order of pre- and post-testing allowing this comparison to be made(p127). However, significant differences were found between the means of PSA and PSB, within sub-groups 4 and 8 ($p<0.05$ in both cases). These results could be attributable to treatment effects for conditions X_1 and X_5. In contrast, the non-significant results associated with PSA and PSB within sub-groups 5, 6 and 7 indicated an absence of any treatment effects for conditions X_2, X_3 and X_4 respectively.

Table 7.5. T-test of response score differences for each sub-group between PSA and PSB.

Sg	N=	Tr	<-- PSA --> Mean	StDv	Tr	<-- PSB --> Mean	StDv	T-test T	sig
1	15	X_n	3.20	0.56	X_o	3.73	0.80	-1.12	ns
2	15	X_o*	3.60	0.73	X_n	3.87	0.52	-1.07	ns
3	17	X_n	3.65	1.00	X_1	3.12	1.22	-1.38	ns
4	17	X_1*	3.94	0.83	X_n	3.29	1.26	1.78	.05
5	11	X_n	2.91	0.83	X_2	3.18	1.08	0.58	ns
6	13	X_n	3.46	0.78	X_3	3.23	0.83	-0.82	ns
7	13	X_n	3.54	0.88	X_4	3.38	0.96	-0.37	ns
8	11	X_n	4.00	0.77	X_5	3.45	0.69	-1.94	.05

(* indicates post-test)

ONEWAY (analysis of variance)

To complement the series of t-tests applied between PSA and PSB, a series of ONEWAY analyses of variance were undertaken on

the within sub-group means (Table 7.6). These were intended to provide evidence of any 'treatment effects' (conditions X_1 to X_5) present as a consequence of In-Service Training undertaken prior to post-testing.

Table 7.6. Oneway ANOVA for response score mean differences across PSA and PSB.

Sub - group	<------- PSA -------> N=	Tr	Mean	StDv	95% Confidence Interval	Sub-gp pair
1	15	X_n	3.20	0.56	2.89 to 3.51	n/s
2	15	X_o	3.60	0.73	3.19 to 4.01	n/s
3	17	X_n	3.65	1.00	3.13 to 4.16	n/s
4	17	X_1	3.94	0.83	3.52 to 4.37	S/gp 5
5	11	X_n	2.91	0.83	2.35 to 3.47	n/s
6	13	X_n	3.46	0.78	2.99 to 3.93	n/s
7	13	X_n	3.54	0.88	3.01 to 4.07	n/s
8	11	X_n	4.00	0.77	3.48 to 4.52	S/gp 5
Sub - group	<------- PSB -------> N=	Tr	Mean	StDv	95% Confidence Interval	Sub-gp pair
1	15	X_o	3.73	0.80	3.29 to 4.18	n/s
2	15	X_n	3.87	0.52	3.58 to 4.15	n/s
3	17	X_1	3.12	1.22	2.49 to 3.74	n/s
4	17	X_n	3.29	1.26	2.64 to 3.94	n/s
5	11	X_2	3.18	1.08	2.46 to 3.91	n/s
6	13	X_3	3.23	0.83	2.73 to 3.73	n/s
7	13	X_4	3.38	0.96	2.80 to 3.97	n/s
8	11	X_5	3.45	0.69	2.99 to 3.92	n/s

Significant differences were indicated between sub-groups 4 and 5 for PSA ($p<0.05$). This result was expected due to the non-equivalent treatment condition assignments for the two sub-groups. However, a significant difference was also registered for the sub-groups 5 and 8 pairing ($p<0.05$). Both latter sub-groups having identical non-treatment condition assignments (X_n). The analysis for PSB indicated an absence of any

significant pairings. This, in-spite of several differing treatment condition assignment across the range of sub-groups.

The results of Tables 7.5 and 7.6 collectively indicate two findings. Firstly, there is evidence to support a combined treatment effect of conditions X_2 and X_5. This is shown from an inspection of the mean difference values for $PSA_{5\&8}$ in conjunction with $PSB_{5\&8}$. In the absence of additional evidence for the X_2 condition, the effect is probably due to X_5. Secondly, there is evidence to support a combined treatment effect of conditions X_1 and X_2. This is shown from an inspection of the mean difference values for $PSA_{4\&5}$ in conjunction with $PSB_{4\&5}$. Again in the absence of further evidence to support the X_2 treatment effect the X_1 condition is probably responsible for the observed differences. Without additional significant differences available within the analysis data for both Tables 7.5 and 7.6, it is difficult to isolate any individual treatment effects.

In summary, the preliminary analysis of the data tended to confirm the expectation that a uni-dimensional analysis would be of limited value. The range of facility values associated with PSA and PSB indicated certain profiles possessed low levels of discrimination (undesirable in a conventional test-item set). However, significant mean score differences were apparent between sub-groups for the same profile-set (Table

7.5). Similarly, significant mean score differences were apparent between profile-sets (Table 7.6). It was anticipated that the multi-dimensional analysis would yield results consistent with these uni-dimensional analysis findings.

The multi-dimensional statistical techniques utilised were grouped into four distinct levels of analysis: (i) within sub-groups: logistic regression; (ii) between sub-groups: inter-correlations and ordering theoretic modelling; (iii) between sub-groups: multiple regression and; (iv) between subjects: homogeneity. The two remaining research aims were analysed using this multi-level framework. Hypotheses were tested through individual or combinations of analyses. The SPSS PC+ statistical program performed the computational components of these analyses (except the ordering theoretic modelling which was calculated with a Microsoft WORKS spreadsheet application).

7.3 Identification of a cognitive simplification strategy.

Aim 1. To investigate the concept of teachers' professional judgements by considering the effects of cognitive simplification strategies on the rating process.

Hypothesis 1. **Teachers' professional judgments are schema-based relying on cognitive simplification strategies which involve systematic rating policy errors.**

Hypothesis 2. **Teachers employ heuristic strategies in the selection and interpretation of student cue information during the rating process.**

Hypothesis 3. **Teachers' professional judgements are based on rating policies which are homogenous.**

The Grantley cohort was identified for the first aim and involved sub-groups 1 and 2. The testing arrangement utilized a paired variation on a post-test only control group design (Campbell and Stanley, 1966). Random assignment to separate treatment conditions ensured pre-treatment equality of sub-groups.

$$\begin{array}{lll} R & & O_{1a} \\ R & X_O & O_{2a} \end{array} \quad \& \quad \begin{array}{lll} R & X_O & O_{1b} \\ R & & O_{2b} \end{array}$$

The depiction O_{1a}, for example, represents the observed ratings for hypothetical profile set A, given by the 1st sub-group sample. The condition X_O was designated as an identity treatment which would represent the equivalent of a 'learning effect' gained during the completion of the pre-test. The first aim was sub-divided into three distinct components; each explored through separate hypotheses.

7.3.1 Within Sub-Groups: logistic regression analysis.

Procedure

The logistic regression analysis utilised the teacher as the unit of analysis. Teacher ratings (dependent variable) were regressed onto the three underlying dimensions (predictor variables) associated with each of the five profiles. This technique is specifically designed for use with dichotomous data and was therefore adopted in preference to that of multiple-regression. A forced entry approach entered all three predictor variables into the regression equation. The program applied an iterative process enabling the formulation of a series of regression weights which optimised the fit of the equation to the data. The iterative process terminated at a log-likelihood default value (0.01%) or when a perfect fit was detected. The correlation between the combined predictor variables with the dependent variable was indicated through a *goodness of fit* value (expressed as a percentage). Beta weights (unstandardised), illustrated within the regression equation, indicated which particular predictor variables maximised this *goodness-of-fit* correlation. T-values determined the significance level of the *independent* contribution made by each of these variable. In practice the *goodness-of-fit* values were, with very few exceptions, 100%. This precluded any consideration of significance attributable

to any regression weight. In these circumstances regression weights were deemed to represent a measure of predictor variable influence rather than significance. In particular, the magnitude and direction of this influence became a direct measure of judgement policy characteristics.

Results

From the regression of the three profile dimension variables (CONGRUENCE status, THRESHOLD and MAXIMUM proficiency) onto each teacher's ratings, within both sub-groups of the Grantley cohort, individual judgement policy equations were generated and are illustrated within Tables 7.7 and 7.8. Since all the profile dimensions were coded zero or one the standard deviations of these independent variables were identical and permitted a direct comparison of the unstandardised regression coefficients (or weights). The mutual independence of the congruence and proficiency dimensions further simplified the interpretation of these weights. A preliminary inspection of both sub-groups across the two profile sets revealed that of the 60 listed regression equations all but one represented perfect-fit solutions. Additional detailed visual inspections were then undertaken for each profile-set.

Table 7.7 Regression of subjects' ratings on the underlying profile dimensions (unstandardised coefficients). (N = 15)

Sub-ject	<------ PSA_1 ------> CON	THR	MAX	GF	<------ PSB_1 ------> CON	THR	MAX	GF
1	0.0	0.0	0.0	1.0	0.0	38.4	0.0	1.0
2	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
3	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
4	38.4	0.0	0.0	1.0	-38.2	77.5	-39.3	1.0
5	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
6	-38.2	77.5	-39.3	1.0	0.0	38.4	0.0	1.0
7	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
8	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
9	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
10	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
11	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
12	38.2	-37.1	-1.1	1.0	0.0	38.4	0.0	1.0
13	38.4	0.0	0.0	1.0	-38.2	1.1	-39.3	1.0
14	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
15	0.0	38.4	0.0	1.0	0.0	38.4	0.0	1.0
Mean	28.1	5.8	17.1	1.0	-5.1	38.5	-5.2	1.0
StDv	21.9	23.6	23.3	0.0	13.0	14.0	13.4	0.0

Of the 15 teachers represented within the PSA_1 regression weight data displayed within Table 7.7, 13 were influenced by the CONGRUENCE status of the student hypothetical profiles. However, the THRESHOLD proficiency cue, similarly depicted within the student profiles, registered an influence for only 3 of the teachers. In contrast, the MAXIMUM proficiency cue was taken into consideration within the decision making process by 9 of those teachers represented. A similar inspection of the PSA_2 regression weight data (Table 7.8), illustrated that of the 15 listed regression equations CONGRUENCE status appeared to be influential in the decisions undertaken by 13 of the teachers. In contrast, both the THRESHOLD and MAXIMUM

proficiency cues illustrated within the student profiles displayed only minimal influence in the case of 11 teachers. The general judgement policy for $PSA_{1\&2}$ across the cohort appeared to be represented by a dominant influence of CONGRUENCE status. Although the influence of the MAXIMUM proficiency cue depicted within the hypothetical profiles had some support this was only partial, restricted primarily to a small cluster of teachers within PSA_1.

Table 7.7 illustrates the PSB_1 regression weight data, which indicates that for all of the 15 teachers involved, the CONGRUENCE status of the student profiles registered only a minimal or negative influence within the decision making process. The THRESHOLD proficiency cue, however, indicated influence in 14 of the total possible judgement policies. The MAXIMUM proficiency cue paralleled that of the CONGRUENCE status pattern, showing minimal or negative influence again in all decision strategies. A similar inspection of the PSB_2 regression weight data, depicted within Table 7.8, indicates the CONGRUENCE status of the profiles registered minimal or negative influence in the decisions of 14 of the 15 teachers. The THRESHOLD proficiency cue, however, appeared to be influential in all 15 instances. Unlike the congruence and proficiency dimensions, which are independent, THRESHOLD and MAXIMUM proficiency cues are not mutually exclusive. Hence, the inter-dependence of the depicted influence of the THRESHOLD and

MAXIMUM proficiency cues is reflected in the latter's minimal weighting values. The overall judgement policy for $PSB_{1\&2}$ across the cohort appeared to be represented by a singular and dominant influence of the THRESHOLD proficiency cue depicted within the profiles.

Table 7.8 Regression of subjects' ratings on the underlying profile dimensions (unstandardised coefficients). (N = 15)

Sub-ject	<------- PSA_2 -------> CON	THR	MAX	GF	<------- PSB_2 -------> CON	THR	MAX	GF
16	0.0	0.0	0.0	1.0	0.0	38.4	0.0	1.0
17	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
18	38.2	1.1	-39.3	1.0	38.2	39.3	-1.1	1.0
19	0.0	9.2	-9.2	0.1	0.0	38.4	0.0	1.0
20	-38.2	39.3	-1.1	1.0	-38.2	39.3	-1.1	1.0
21	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
22	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
23	38.4	0.0	0.0	1.0	-38.2	39.3	-1.1	1.0
24	38.2	39.3	-1.1	1.0	0.0	38.4	0.0	1.0
25	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
26	38.2	-37.1	-1.1	1.0	0.0	38.4	0.0	1.0
27	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
28	38.2	1.1	37.1	1.0	0.0	38.4	0.0	1.0
29	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
30	38.4	0.0	0.0	1.0	-38.2	39.3	-1.1	1.0
Mean	28.1	3.6	1.5	0.9	-5.1	38.6	-0.3	1.0
StDv	22.0	17.0	17.0	0.2	19.1	0.4	0.5	0.0

Comparison between PSA and PSB (Tables 7.7 & 7.8) across the Grantley cohort indicated the adopted judgement policies tended to be complementary. The use of predominantly mutually exclusive judgement policies across the two profile sets was apparent on an individual teacher and overall sub-group level. The latter represented by mean judgement policies for each

profile-set. Noticeably, negative regression weights were evident across all three dimensions of CONGRUENCE status, THRESHOLD and MAXIMUM proficiency. The negative value of the regression weight MAX (for instance subject 18, PSA_2 within Table 7.8) indicates, in practice, that faced with a profile depicting maximum proficiency the teacher would assign a zero rating (ie. failure of the profile to match the given criterion). Similarly, the negative value of the regression weight CON (for example subject 23, PSB_2 within Table 7.8) indicates the teacher would, if faced with a profile depicting congruence, assign a zero rating. Finally, any substantial differences between the pairings of $PSA_{1\&2}$ or $PSB_{1\&2}$ were not evident from the visual inspections of the judgement policy regression equations. Hence, the determination of a 'learning effect' associated with the X_o treatment condition was inconclusive.

7.3.2 Between Sub-Groups: correlational/hierarchical analysis.

Procedure (correlational)

The correlational procedure utilised individual subjects' underlying dimensions and profiles as the units of analysis. This involved two distinct aspects. The first produced a series of inter-regression weight correlations, utilizing the beta coefficients calculated from logistic regression analysis. Correlational matrices were calculated for each profile-set,

and across the range of sub-groups. The second aspect considered a series of inter-profile correlations. Correlational matrices were calculated for each profile set, and across the range of sub-groups. In both cases significant correlations (ie. $p<0.05$ and $p<0.01$) were identified within each matrix.

Results (correlational)

Inter-regression weight correlations were calculated for both sub-groups of the cohort and are shown within Table 7.9. Correlation values enabled any relationship between the independent congruence and proficiency dimensions to be determined. Additionally, the association between the dependent proficiency cue variables THRESHOLD and MAXIMUM was available for examination.

Table 7.9 Correlational Coefficients associated with the unstandardised regression values for PSA & PSB. (N=15)

<--------- PSA_1 --------->				<--------- PSB_1 --------->			
	CON	THR	MAX		CON	THR	MAX
CON	--	-61**	13	CON	--	00	100**
THR		--	31	THR		--	00
MAX			--	MAX			--
<--------- PSA_2 --------->				<--------- PSB_2 --------->			
	CON	THR	MAX		CON	THR	MAX
CON	--	-61**	41	CON	--	-50	50
THR		--	-28	THR		--	-100**
MAX			--	MAX			--

* $p<.05$ ** $p<.01$

When viewed collectively the sub-group correlations across both profile sets indicated a consistent pattern of results across PSA but not PSB. CONGRUENCE and THRESHOLD proficiency demonstrated a negative correlation ($p<0.01$) within PSA_1 and PSA_2. CONGRUENCE and MAXIMUM proficiency registered a positive correlation ($p<0.05$) within PSB_1. Finally, THRESHOLD and MAXIMUM proficiency depicted a negative correlation; achieving significance ($p<0.01$) within PSB_2. The consistency across PSA indicates two important features. Firstly, the possible existence of a definite relationship between CONGRUENCE status and proficiency; for the THRESHOLD proficiency cue. Secondly, an underlying similarity of judgement policy functioning for teachers across PSA in both pre- and post-test conditions. In contrast, the inconsistency across PSB indicated an underlying difference in judgement policy in pre- and post-test conditions.

Inter-profile correlations were calculated for both sub-groups of the cohort (Table 7.10). This allowed the relationship between the individual profiles within each set to be considered. The adoption of complementary judgement policies would be available for inspection across PSA and PSB on an individual profile level.

Table 7.10 Correlational Coefficients associated with (N=15) the profile facilities across PSA & PSB.

<--------- PSA_1 --------->						<--------- PSB_1 --------->					
	E	S	J	Q	D		H	A	T	O	K
E	--	-24	-21	16	.	H	--	.	.	.	.
S		--	15	-68*	.	A		--	100**	68*	.
J			--	-10	.	T			--	68*	.
Q				--	.	O				--	.
D					--	K					--

<--------- PSA_2 --------->						<--------- PSB_2 --------->					
	E	S	J	Q	D		H	A	T	O	K
E	--	-08	31	-16	.	H	--	.	13	.	.
S		--	-45	13	.	A		--	.	.	.
J			--	-16	.	T			--	.	.
Q				--	.	O				--	.
D					--	K					--

* $p<.05$ ** $p<.01$

Inspection of the sub-group correlations across both profile-sets indicated no consistent patterns. Although specific and individual profile-pairs achieved significance, these appeared to be unrelated to any other corresponding results. For instance, within PSA_1 profiles S and Q depicted a significant negative correlation ($p<0.05$); a result not reflected within PSA_2. Similarly, within PSB_1, for example, profiles A and T registered a significant positive correlation ($p<0.01$); but not within PSB_2, this result was not computable. The occurrence of non-computable coefficients indicated the limitations of the small sub-group sizes and the data type.

Procedure (hierarchical)

The hierarchical analysis was based on an ordering theoretic technique (Bart and Krus, 1973). These were undertaken for each profile set, across the range of samples. To establish pre-requisite relationships, comparisons were made of all profile pair combinations. Disconfirmatory matrices were calculated using the *adjusted* response data. These matrices illustrated the number of pre-requisite violations for each profile-pair. Hierarchical diagrams were constructed from these matrices. These illustrated the prerequisites relationships for pairs of profiles. In each case a tolerance level of 10% was arbitrarily established. This value allowed only one violation per pre-requisite relationship to occur before the hierarchical pairing was considered untenable. Validation measures, such as the coefficient of scalability, reproducibility and percentage improvement were not calculated.

Results (hierarchical)

To facilitate a hierarchical analysis of the individual profile response data disconfirmatory matrices were calculated for each sub-group of the Grantley cohort (Table 7.11). 'Ordering-theoretic' hierarchy diagrams were constructed for each matrix with a tolerance level of 10% (Fig 7.1 & 7.2). The adoption of complementary judgement policies depicted within the hierarchy structures associated with PSA and PSB on an individual profile level.

Table 7.11 Disconfirmatory response patterns for two-profile prerequisite relation orderings across PSA. (N = 15)

<-- Hierarchy Matrix PSA_1 -->					
	E	S	J	Q	D
E	-	13	67	67	73
S	27	-	73	87	87
J	7	0	-	13	13
Q	0	7	7	-	7
D	0	0	0	0	-

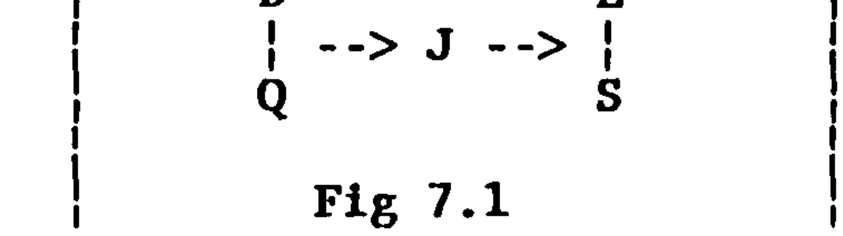


Fig 7.1

<-- Hierarchy Matrix PSA_2 -->					
	E	S	J	Q	D
E	-	7	13	27	27
S	60	-	67	73	80
J	13	13	-	27	27
Q	7	0	7	-	7
D	0	0	0	0	-

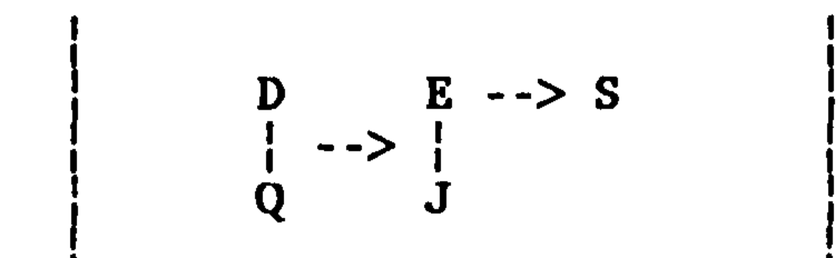


Fig 7.2

The hierarchy patterns for $PSA_{1\&2}$ and $PSB_{1\&2}$ are illustrated by Figures 7.1 & 7.2 and 7.3 & 7.4 within Tables 7.11 and 7.12. These diagrams represent pre-requisite relations. For example D is a pre-requisite of J (figure 7.1). This indicates that teachers correctly responding to J were also successful in their response to D (within the given level of tolerance). This relationship is not necessarily reversible. Figures 7.1, 7.2, 7.3 and 7.4 indicate a degree of structural similarity for each of the two profile pairings (i.e. $PSA_{1\&2}$ and $PSB_{1\&2}$). A four profile correspondence within both pairings was evident. Within $PSA_{1\&2}$ profiles J and S were interchanged this involved profiles T and O for $PSB_{1\&2}$. Although both hierarchical

structures depict multi-dimensional pre-requisite relationships these differ noticeably between the two profile-sets. In contrast to the results of the inter-regression and inter-profile correlations the constructed hierarchies appear to be stable in-spite of the small sample sizes involved. In fact the absence of contradictory hierarchy patterns (utilizing a 10% tolerance level) support the validity of the emergent pre-requisite relations involved. However, with two groupings of only five profiles any further hierarchical analysis was not possible.

Table 7.12 Disconfirmatory response patterns for two-profile prerequisite relation orderings across PSB. (N = 15)

<-- Hierarchy Matrix PSB_1 -->

	H	A	T	O	K
H	-	93	93	87	100
A	0	-	0	0	7
T	0	0	-	0	7
O	0	7	7	-	13
K	0	0	0	0	-

<---- PSB_1 Hierarchy ---->

A
|
K --> O --> H
|
T

Fig 7.3

<-- Hierarchy Matrix PSB_2 -->

	H	A	T	O	K
H	-	93	73	93	93
A	0	-	0	0	0
T	0	20	-	20	20
O	0	0	0	-	0
K	0	0	0	0	-

<---- PSB_2 Hierarchy ---->

A
|
K --> T --> H
|
O

Fig 7.4

7.3.3 Between Sub-Groups: multiple-regression analysis.

Procedure

The multiple regression technique utilised each sample as the unit of analysis. Beta weights (from the logistic regression procedure) were regressed onto the Assessment Profile and Teacher Biographic Information respectively. The Assessment Profile ratings were considered to explicitly represent the judgement policy of each teacher. Hence, these ratings (predictor variables) were expected to correlate with the regression weights (dependent variables). The teacher biographic details were considered to have the potential to explain judgement policy differences. Therefore, these details (moderator variables) were similarly expected to correlate with the regression weights (dependent variables).

In both procedures a forced entry approach entered all the variables into the regression equation. The program applied an iterative process allowing the formulation of a series of regression weights which optimised the fit of the equation to the data. The correlation between the combined predictor variables with the dependent variable was indicated through a multiple R value (expressed as an R^2 value). Beta weights (standardised), depicted within the regression equation, established which particular predictor variables maximised the multiple correlation. T-values indicated the significance level

of the independent contribution made by each of these variables taking account of all other predictors. Previous research (Slovic and Lichtenstein, 1971) has indicated that R values of 0.70 and above represents a good level of multiple correlation for policy capturing techniques. Therefore, within this study R^2 values of 0.50 were considered to indicate substantial predictor variable influence. Additionally, R^2 values of 0.30 and 0.70 were considered to indicate moderate and very substantial predictor variable influence, respectively.

Results

The regression of Assessment Profile variables onto individual teacher regression equations (logistic regression weights), within both sub-groups of the Grantley cohort generated 'policy to practice' correlations (Table 7.13 and 7.14). Unlike the profile dimensions utilised within the intial regression analysis, the Assessment Profile information was not coded one or zero. Hence, standardised regression weights were adopted for the purpose of analysis. A preliminary inspection of both sub-groups across the two profile sets revealed a moderate or substantial set of combined predictor variable correlations on 9 of 12 possible occasions. Additional inspections were then undertaken for each profile set.

Table 7.13 Multiple Regression of Assessment Profile Information onto teachers' regression equations (standardised coefficients). Sgp 1 (N = 15)						
	<------- PSA_1 ------->			<-------- PSB_1 ------>		
API	CON	THR	MAX	CON	THR	MAX
PDF	0.43	-0.79	0.73	0.28	0.36	0.28
NDF	-0.02	0.46	-0.12	-0.26	0.21	-0.26
PSC	0.23	-0.29	0.51	0.14	0.10	0.14
WKS	0.37	-0.33	0.36	0.13	-0.49	0.13
CNG	0.24	0.03	0.04	-0.46	0.33	-0.46
R^2	0.34	0.24	0.52	0.19	0.40	0.19

* $p<.05$ ** $p<.01$

Table 7.13 illustrates the PSA_1 multiple-regression weight data. The results indicated a moderate correlation (R^2=0.34) of the combined predictor variables with the CONGRUENCE status component of teachers' judgement policies. A substantial correlation (R^2=0.52) was registered with the combined predictor variables and the MAXIMUM proficiency component. In both cases no individual variables made significant independent contributions. Inspection of PSA_2 (Table 7.14) indicated a substantial correlation with the combined predictor variables and both the CONGRUENCE (R^2=0.88) and THRESHOLD proficiency (R^2=0.76) components of judgement policy. Additionally, a moderate correlation was registered for the MAXIMUM proficiency (R^2=0.36) component of the policy. The individual Proficiency Score variable (PSC) correlated significantly ($p<0.05$) with the CONGRUENCE component of the policy. This represented a covariation in congruence influence and a teacher's perception of proficiency. Similarly the individual Short-Term-Recall

variable (WKS) correlated significantly ($p<0.05$) with the THRESHOLD proficency component of the policy. This identified a covariation in proficiency influence and a subject's perception of short-term-recall functioning. The negative value shows greater proficiency influence was associated with the diminished importance attached to short-term-recall effect.

Table 7.14 Multiple Regression of Assessment Profile Sgp 2 (N = 15) Information onto teachers' regression equations (standardised coefficients).

API	<------- PSA_2 -------> CON	THR	MAX	<------- PSB_2 -------> CON	THR	MAX
PDF	-0.90	-0.63	-1.48	0.78	-0.78	0.78
NDF	0.77	0.72	1.01	-0.10	0.10	-0.10
PSC	0.46*	-0.63	-0.40	0.16	-0.16	0.16
WKS	0.30	-0.77*	-0.05	0.88*	-0.88*	0.88*
CNG	0.25	0.05	-0.45	0.20	-0.22	0.22
R^2	0.88	0.76	0.36	0.59	0.59	0.59

* $p<.05$ ** $p<.01$

The PSB_1 data, displayed within Table 7.13, revealed a moderate correlation (R^2=0.40) of the combined predictor variables and the THRESHOLD proficiency component of the judgement policy. No individual variables made a significant independent contribution. A similar inspection of PSB_2 (Table 7.14) indicated a substantial correlation (R^2=0.59 in all three cases) with the combined predictor variables and the CONGRUENCE status, THRESHOLD and MAXIMUM proficiency components of the judgement policy. Further to this, the individual short term recall variable (WKS) correlated significantly ($p<0.05$) with

all three components of the judgement policy. While the CONGRUENCE and MAXIMUM proficiency components registered positive correlations, THRESHOLD proficiency covaried negatively. These identified a covariation in judgement policy and a subject's perception of short term recall functioning. On this occasion, individual component influence was associated with both positive and negative correlations attached to short-term-recall effects.

Consideration of PSA and PSB (Tables 7.13 & 7.14) across the two sub-groups of the Grantley cohort indicated a potential heuristic decision strategy was in operation. The strategy, in effect, was thought to be responsible for generating the differences between expressed policy and actual practice. Initially, this was evident through the complementary judgement policies adopted between the two profile sets (consistent across both sub-groups). More specifically, teachers perception of the utility of Statements of Attainment as assessment criteria (depicted through the CNG variable) appeared to have little association with the actual influence of the CONGRUENCE status of the hypothetical student profiles. In addition, the proficiency ratings of teachers (registered through the PSC variable) provided a generally consistent pattern of association with the CONGRUENCE status component of judgement policy. A similar consistent pattern was not evident, however, for the THRESHOLD and MAXIMUM proficiency components. Finally, teachers' perception of short-term-recall compensation

(represented through the WKS variable) demonstrated a generally consistent pattern of association with all three components of judgement policy. This latter result highlighted the extensive relationship of explicit policy with implicit practice.

Teachers' biographic variables were regressed onto the judgement equations (logistic regression weights), within both sub-groups of the Grantley cohort, and the resultant individual moderator correlations are shown within Tables 7.15 and 7.16. In common with the Assessment Profile analysis, standardised regression weights were adopted for the purposes of analysis. An initial inspection of both sub-groups across the two profile sets revealed a moderate or substantial set of combined predictor variable correlations in 6 of the possible 12 outcomes (R^2 values). Further, more detailed and comparative inspections were then undertaken across the two profile sets for this sample.

Table 7.15 Multiple Regression of Teacher Biographic Sgp 1 (N = 15) Information onto teachers' regression equations (standardised coefficients).

	<------- PSA_1 ------->			<------- PSB_1 ------->		
TBI	CON	THR	MAX	CON	THR	MAX
YTEA	0.16	-0.15	0.36	0.17	0.16	0.17
PSUB	--	--	--	--	--	--
CREX	0.04	0.30	0.04	-0.24	0.07	-0.24
NCEX	-0.05	-0.04	-0.48	-0.44	0.01	-0.44
R^2	0.02	0.16	0.34	0.27	0.02	0.27

* p<.05 ** p<.01

Represented within Table 7.15 are the PSA_1 and PSA_2 correlation results. The PSA_1 data indicated a moderate correlation (R^2=0.34) existed between the combined predictor variables and the MAXIMUM proficiency component of the judgement policy. No individual variable made a significant independent contribution. A similar inspection of the PSA_2 data (Table 7.16) indicated a moderate correlation (R^2=0.36) with the combined predictor variables and the CONGRUENCE component of judgement policy. Again no individual variable made a significant independent contribution. A substantial correlation (R^2=0.66) was found between the combined predictor variables and the MAXIMUM proficiency component. Teachers' principal subject variable (PSUB) indicated a significant ($p<0.05$) correlation with the MAXIMUM proficiency component of the policy. This identified a negative covariation between Mathematics as a main taught subject and MAXIMUM proficiency influence. That is, the MAXIMUM proficiency cue had a greater influence over the decision making process of non-mathematicians than it did for the subject specialist.

The PSB_1 data, represented within Table 7.15, revealed no significant correlations with the combined predictor variables and any of the components of judgement policy. A similar inspection of the PSB_2 data (Table 7.16) indicated a substantial correlation (R^2=0.33 in all three cases) with the combined predictor variables and all three components of

judgement policy. However, no significant individual variable contributions were evident across these components.

Table 7.16 Multiple Regression of Teacher Biographic Sgp 2 (N = 15) Information onto teachers' regression equations (standardised coefficients).

TBI	<------- PSA_2 -------> CON	THR	MAX	<------- PSB_2 -------> CON	THR	MAX
YTEA	-0.39	-0.11	0.19	-0.52	0.52	-0.52
PSUB	-0.40	0.00	-0.71*	-0.51	0.51	-0.51
CREX	-0.39	0.41	-0.01	-0.10	0.10	-0.10
NCEX	0.46	-0.48	-0.13	0.18	-0.18	0.18
R^2	0.36	0.13	0.66	0.33	0.33	0.33

* p<.05 ** p<.01

A comparison between PSA and PSB (Table 7.15 & 7.16) generally revealed a degree of inconsistency for all four potential moderator variables. This inconsistency created difficulties for the identification of potential moderator variables. Although Mathematics as a principal subject (expressed through PSUB), achieving significance within PSA_2, did indicate a degree of eligibility for this variable. Any meaningful interpretation, however, is restricted in the absence of supportive evidence from PSA_1, PSB_1 and PSB_2.

7.3.4 Between Subjects: homogeneity analysis.

Procedure

The investigation of judgement policy homogeneity utilised the individual subject as the unit of analysis. Judgement policy regression equations were compared using the test for common slopes within the analysis of covariance. If the test indicated the judgement policies were homogeneous then it would be appropriate to report an overall average regression equation for the sub-group. Because the assumption of between-subjects might prove untenable, the variation of individual regression equations was inspected. This enabled groups or clusters of teachers utilizing the same policy to be identified. The occurrence of three or more teachers (an arbitrary criterion) adopting the same policy constituted a cluster.

Results

The regression equation frequency distribution for PSA_1, illustrated within Table 7.17, indicates two predominant key judgement policies were utilised by teachers. The influence of CONGRUENCE status was prominent in both key policies; which accounted for 11 of the 15 teacher judgements in total. Although, for one policy (38.2, 1.1, 37.1)), eight teachers also acknowledged the importance of the MAXIMUM proficiency cue. The remaining four teachers adopted policies which were unique; each emphasising differing influences of CONGRUENCE

status, THRESHOLD and MAXIMUM proficiency. A similar consideration of the regression equation frequency distribution for PSA_2 (Table 7.17) indicated the dominance of one key judgement policy. The influence of CONGRUENCE status was prominent in this policy; which accounted for 7 decision strategies. Two teachers adopted policies which depicted the influence of MAXIMUM proficiency. The remaining six teachers each adopted individual policies; and emphasising differing influences of the CONGRUENCE status, THRESHOLD and MAXIMUM proficiency cues.

An inspection of the regression equation frequency distribution for PSB_1, represented within Table 7.17, indicated only one key judgement policy was utilised by teachers. The influence of the THRESHOLD proficiency cue was dominant within this policy, accounting for 13 judgements in total. The remaining two teachers each adopted policies which were unique and dissimilar. A similar consideration of the frequency distribution for PSB_2 (Table 7.17) revealed the utilisation of two key judgement policies by teachers. Again the influence of the THRESHOLD proficiency cue was predominant across both policies; which accounted for 14 of the 15 teacher judgements in total. Although, in the case of one policy (-38.2, 39.3, -1.1), three teachers also recognized the importance of CONGRUENCE status. The remaining, and therefore unique,

judgement policy depicted a balance between CONGRUENCE status and THRESHOLD proficiency influence.

Table 7.17 Frequency distribution of teachers' regression equations across PSA and PSB (unstandardised coefficients).

Tr	<----- PSA_1 -----> CON	THR	MAX	Fr	T	<----- PSB_1 -----> CON	THR	MAX	F_r
X_n	38.2	1.1	37.1	8	X_o	0.0	38.4	0.0	13
	38.4	0.0	0.0	3		-38.2	1.1	-39.3	1
	38.2	-37.1	-1.1	1		-38.2	77.5	-39.3	1
	0.0	38.4	0.0	1					
	0.0	0.0	0.0	1					
	-38.2	77.5	-39.3	1					
MEAN	28.1	5.8	17.1			-5.1	38.5	-5.2	
Tr	<----- PSA_2 -----> CON	THR	MAX	Fr	T	<----- PSB_2 -----> CON	THR	MAX	Fr
X_o	38.2	39.3	-1.1	1	X_n	38.2	39.3	-1.1	1
	38.2	1.1	37.1	2		-38.2	39.3	-1.1	3
	38.4	0.0	0.0	7		0.0	38.4	0.0	11
	38.2	-37.1	-1.1	1					
	38.2	1.1	-39.3	1					
	0.0	0.0	0.0	1					
	0.0	9.2	-9.2	1					
	-38.2	39.3	-1.1	1					
MEAN	28.1	3.6	1.5			-5.1	38.6	-0.3	

(bold print indicates judgement policy frequency > 2)

Comparison between PSA and PSB across the cohort, revealed an adoption of judgment policies which tended to be complementary. Between both profile-sets the two dominant decision strategies appeared to be mutually exclusive regarding the importance attached by teachers to the dimensional information depicted within the regression equations. Within PSA teachers adopted

decision strategies which indicated a dominance of CONGRUENCE status; although a variability of influence of MAXIMUM proficiency was also discernible. In contrast, PSB revealed teachers' tended to utilise decision strategies with a predominant THRESHOLD proficiency influence. Of the 60 policies illustrated only two depicted the 'perfect solution' of a balanced judgement (38.2, 39.3, -1.1). The between-subjects homogeneity analysis confirmed the acceptability of reporting an overall regression equation for each sub-group; except PSB_2. Finally, any significant difference between the $PSA_{1\&2}$ judgement policy distributions was not evident. However, the significant difference in homogeneity between $PSB_{1\&2}$ is interpretable as a potential 'learning effect' associated with the X_o treatment condition.

7.3.5 Discussion of the findings.

The investigation of 'teacher professional judgement' revealed evidence for the existence of a schema-based cognitive simplification strategy. The results of the logistic regression analysis demonstrated the adoption of two potential schemata; one centred upon proficiency the other focussing on congruence. These mutually exclusive policies, evident at a sub-group level, were frequently demonstrated within the judgements of individual teachers. The absence of a significant 'learning

effect' was shown by the stability of judgement policies within profile-sets across equivalent treatment conditions.

The nature of the decision strategies adopted by teachers was shown by the multiple-regression analysis. Significantly, implicit policy and explicit practice differences were evident. Rather than a redundancy of association between policy and practice, inappropriate relationships were discernible with proficiency policy registering a predictive influence upon congruence practice. In addition, the utility of teacher biographic information to explain decision strategy differences also revealed any practical benefits were restricted to combinations of predictor variables. Finally, the variation of judgement policies between the two profile sets revealed the adoption of an homogenous decision strategies within $PSA_{1\&2}$ but not $PSB_{1\&2}$. This latter result providing evidence of a potential pre-test/post-test 'learning effect'.

7.4 INSET and teachers' professional judgements.

Aim 2. To explore the effect on teacher professional judgement of modifications to rating policies brought about through In-Service Education and Training.

Hypothesis 4. **The professional judgements of teachers are significantly influenced by In-Service Education and Training.**

The York cohort was identified for the fulfilment of this aim and involved sub-groups 3 and 4. The testing arrangement again utilized a paired variation on a post-test only control group design, with random assignment to separate treatment groups. Thus ensuring pre-treatment sub-group equality.

$$\begin{array}{llll} R & & O_{3a} & \\ R & X_O & O_{4a} & \end{array} \quad \& \quad \begin{array}{llll} R & X_O & O_{3b} & \\ R & & O_{4b} & \end{array}$$

The depiction O_{3a}, for instance, represents the observed ratings for hypothetical profile set A, given by the 3rd sub-group sample. The condition X_1 was designated as the In-Service Education and Training treatment condition; and would be delineated as an 'INSET effect' gained prior to the undertaking of the post-test. This second aim was explored through one hypothesis.

7.4.1 Within Sub-Groups: logistic regression analysis.

Results

From the regression of the three profile dimension variables (CONGRUENCE status, THRESHOLD and MAXIMUM proficiency) onto each teachers' ratings, within both sub-groups of the York

cohort, individual judgement policy equations were generated (Tables 7.18 and 7.19). A preliminary inspection of both samples across the two profile sets revealed that of the 64 listed regression equations all but five represented perfect fit solutions. Detailed visual inspections were then undertaken allowing further comparisons to be considered.

Table 7.18 Regression of subjects' ratings on the underlying profile dimensions (unstandardised coefficients). (N = 17)

Sub-ject	<------ PSA_3 ------>				<------ PSB_3 ------>			
	CON	THR	MAX	GF	CON	THR	MAX	GF
31	0.0	0.0	0.0	1.0	0.0	38.4	0.0	1.0
32	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
33	0.0	38.4	0.0	1.0	0.0	38.4	0.0	1.0
34	38.2	1.1	37.1	1.0	-38.2	77.5	-39.3	1.0
35	38.4	0.0	0.0	1.0	0.0	38.4	0.0	1.0
36	*0.0*	*9.2*	*-9.2*	*0.1*	*0.0*	*-9.2*	*9.2*	*0.1*
37	38.4	0.0	0.0	1.0	38.2	-1.1	39.3	1.0
38	0.0	38.4	0.0	1.0	-38.2	77.5	-39.3	1.0
39	38.2	39.3	-1.1	1.0	0.0	38.4	-38.4	1.0
40	38.2	39.3	-1.1	1.0	-38.2	77.5	-39.3	1.0
41	38.2	-37.1	-1.1	1.0	0.0	-38.4	38.4	1.0
42	0.0	38.4	0.0	1.0	38.2	-1.1	39.3	1.0
43	0.0	0.0	-38.4	1.0	-38.2	-1.1	39.3	1.0
44	38.4	0.0	0.0	1.0	*0.0*	*-9.2*	*9.2*	*0.1*
45	*0.0*	*9.2*	*-9.2*	*0.1*	38.2	39.3	-1.1	1.0
46	38.2	1.1	-39.3	1.0	-38.2	77.5	-39.3	1.0
47	38.2	39.3	-1.1	1.0	-38.2	77.5	-39.3	1.0
Mean	22.5	14.5	-3.2	0.9	-6.7	32.9	-3.6	0.9
StDv	18.8	22.3	15.9	0.2	27.0	36.2	30.0	0.3

Inspection of the PSA_3 regression weight data, Table 7.18, indicated that CONGRUENCE status of the student profiles influenced the decision making process of 10 of the 17 teachers. In contrast, MAXIMUM proficiency depicted minimal

influence within the decision making process for 13 teachers. Only three teachers (subjects 39, 40 & 47) demonstrated a balanced judgement policy of CONGRUENCE status and THRESHOLD proficiency influence within this profile set. A similar consideration of the PSA_4 regression weight data, represented within Table 7.19, illustrated that of the 17 judgements, CONGRUENCE status appeared to be influential in the decisions undertaken by 15 of the teachers. In contrast, the MAXIMUM proficiency cue, depicted within the student profiles, displayed a minimal or negative influence in the case of 14 teachers. Five teachers (subjects 48, 51, 53, 58 & 60) demonstrated a balanced judgement policy of CONGRUENCE status and THRESHOLD proficiency influence on this occasion. The overall judgement policy for PSA_{364} across the cohort was represented by a dominant influence of CONGRUENCE status. Although the negative influence of the MAXIMUM proficiency cue had a degree of support, this was only partial and restricted to a number of teachers (subjects 52, 55, 61, 62 & 64) within PSA_4.

Of the 17 judgements represented within the PSB_3 regression weight data (Table 7.18), the THRESHOLD proficiency cue was influential in the decision making process of 11 teachers. In contrast, no similar patterns were discernible for the influence of CONGRUENCE status or MAXIMUM proficiency; substantial variability was evident in both aspects. Inspection of the PSB_4 regression weight data, detailed within Table 7.19,

illustrated that of the 17 listed judgement policies the THRESHOLD proficiency cue appeared to be influential in the decisions taken by 13 of the teachers. Whereas the MAXIMUM proficiency cue displayed minimal influence in the case of 13 teachers. The general judgement policy for $PSB_{3\&4}$ across the cohort appeared to be represented by a dominant influence of the THRESHOLD proficiency cues depicted within the student profiles. The negative influence of the MAXIMUM proficiency cue had some degree of support, however, this was mainly restricted to a group of teachers (subjects 34, 38, 39, 40, 46 & 47) within PSB_3.

Table 7.19 Regression of subjects' ratings on the underlying profile dimensions (N = 17) (unstandardised coefficients).

Sub-ject	<------ PSA_4 ------> CON	THR	MAX	GF	<------ PSB_4 ------> CON	THR	MAX	GF
48	38.2	39.3	-1.1	1.0	0.0	38.4	0.0	1.0
49	38.4	0.0	0.0	1.0	0.0	-38.4	38.4	1.0
50	38.2	-1.1	39.3	1.0	-38.2	77.5	-39.3	1.0
51	38.2	39.3	-1.1	1.0	-38.2	39.3	-1.1	1.0
52	0.0	38.4	-38.4	1.0	-38.2	-1.1	39.3	1.0
53	38.2	39.3	-1.1	1.0	-38.2	39.3	-1.1	1.0
54	38.2	-37.1	-1.1	1.0	0.0	38.4	0.0	1.0
55	38.2	1.1	-39.3	1.0	38.2	39.3	-1.1	1.0
56	38.2	-37.1	-1.1	1.0	0.0	0.0	0.0	1.0
57	*0.0*	*9.2*	*-9.2*	*0.1*	0.0	38.4	0.0	1.0
58	38.2	39.3	-1.1	1.0	-38.2	39.3	-1.1	1.0
59	38.2	-1.1	39.3	1.0	38.2	39.3	-1.1	1.0
60	38.2	39.3	-1.1	1.0	38.2	39.3	-1.1	1.0
61	38.2	1.1	-39.3	1.0	-38.2	39.3	-1.1	1.0
62	38.2	1.1	-39.3	1.0	0.0	0.0	0.0	1.0
63	38.4	0.0	0.0	1.0	*0.0*	*9.2*	*-9.2*	*0.1*
64	38.2	1.1	-39.3	1.0	38.2	39.3	-1.1	1.0
Mean	33.7	10.1	-7.9	0.9	-4.5	28.0	1.2	0.9
StDv	13.4	28.0	20.8	0.0	18.9	34.2	25.0	0.0

Across the York cohort (Tables 7.18 & 7.19) the adopted judgement policies appeared to be complementary. As with the Grantley cohort, the use of predominantly mutually exclusive judgement policies across the two profile sets was evident both at an individual teacher and overall sample level. Representation of the latter was through mean judgement policy values documented for each profile-set. A potential 'INSET effect' was discernible from the comparison of sample judgement policies in pre- and post-treatment (X_1) conditions. For example, across $PSA_{3\&4}$ (Tables 7.18 & 7.19), an increase in the positive influence of CONGRUENCE status and negative influence of the MAXIMUM proficiency cue was noticeable after the X_1 treatment. Similarly, across $PSB_{3\&4}$, an increase in the positive influence of the THRESHOLD proficiency cue and negative influence of MAXIMUM proficiency was evident within the X_1 condition. However, the significance of such differences could not be easily established from the visual inspection procedure alone. Hence, the determination of a potential 'INSET effect' associated with the X_1 treatment condition was inconclusive.

7.4.2 Between Sub-Groups: correlational/hierarchical analysis.

Results (correlational)

Inter-regression weight correlations were calculated for both sub-groups of the cohort, and are shown within Table 7.20. These values allowed the relationship between the independent congruence and proficiency dimensions to be examined in the context of a potential 'INSET effect'. Further to this, the association between the dependent THRESHOLD and MAXIMUM proficiency variables was available for examination.

Table 7.20 Correlational Coefficients associated with the unstandardised regression values for PSA & PSB. (N=17)

<--------- PSA_3 --------->				<--------- PSB_3 --------->			
	CON	THR	MAX		CON	THR	MAX
CON	--	-35	13	CON	--	-57*	62**
THR		--	-13	THR		--	-92**
MAX			--	MAX			--

<--------- PSA_4 --------->				<--------- PSB_4 --------->			
	CON	THR.	MAX		CON	THR	MAX
CON	--	-30	46	CON	--	-06	03
THR		--	-29	THR		--	-80**
MAX			--	MAX			--

* $p<.05$ ** $p<.01$

An inspection of the two sample correlations across both profile-sets collectively indicated a relatively consistent pattern of results. CONGRUENCE and THRESHOLD proficiency registered a negative correlation ($p<0.05$) within PSB_3.

CONGRUENCE and MAXIMUM proficiency registered a positive correlation ($p<0.01$) within PSB_3. Finally, THRESHOLD and MAXIMUM proficiency demonstrated a negative correlation ($p<0.01$) on two occasions within PSB_3 and PSB_4. The inconsistency of results across PSB indicates two important features. Firstly, a potential INSET effect evident with the correlations pairs of CON & THR and CON & MAX. In both instances the significance is reversed in sign between the two treatment conditions (X_1 and X_n). Secondly, the significance of the relationship between THR and MAX is maintained across the two treatment conditions (X_1 and X_n).

Table 7.21 Correlational Coefficients associated with (N=17) the profile facilities across PSA & PSB.

<--------- PSA_3 --------->					
	E	S	J	Q	D
E	--	-27	-07	-07	.
S		--	38	38	.
J			--	19	.
Q				--	.
D					--

<--------- PSB_3 --------->					
	H	A	T	O	K
H	--	26	20	25	-17
A		--	31	18	63**
T			--	-34	49
O				--	-20
K					--

<--------- PSA_4 --------->					
	E	S	J	Q	D
E	--	.	.	.	.
S		--	18	42	-34
J			--	-30	-17
Q				--	-24
D					--

<--------- PSB_4 --------->					
	H	A	T	O	K
H	--	31	52	20	20
A		--	31	23	23
T			--	-34	39
O				--	-13
K					--

* $p<.05$ ** $p<.01$

Inter-profile correlations were calculated for both sub-groups of the cohort and are depicted within Table 7.21. These values enabled the relationship between the individual student

profiles within each set to be viewed in the context of the potential 'INSET effect'. The adoption of complementary judgement policies would be available for more detailed inspection across PSA and PSB on an individual profile level.

A collective examination of the sample correlations across both profile sets indicated no consistent patterns. On this occasion only one specific profile pair achieved significance. This positive correlation ($p<0.01$) occurred within PSB_3 and involved the profile pairing of A and K; a result not repeated within PSB_4. As with the Grantley cohort the occurrence of non-computable coefficients indicated the limitations associated with the small sub-group sizes of the York cohort. Although, the frequency of occurrence of these was noticeably less in comparison to the Grantley samples; and in the case of PSB, non-computable coefficients were absent altogether.

Results (hierarchical)

The hierarchical analysis of the individual hypothetical profile response data for the York cohort required the calculation of disconfirmatory matrices for each sub-group and profile-set (Table 7.22). 'Ordering Theoretic' hierarchy diagrams were then constructed for each matrix with a tolerance level of 10% (Fig. 7.5 & 7.6). The adoption of differing, if not complementary, judgement policies would be represented within the hierarchy structures for PSA and PSB. This would be

available in pre- and post-treatment conditions, on an individual profile level.

Table 7.22 Disconfirmatory response patterns for two-profile prerequisite relation orderings across PSA. (N = 17)

<--- Hierarchy Matrix PSA_3 --->

	E	S	J	Q	D
E	-	24	35	35	41
S	41	-	41	41	59
J	12	0	-	12	18
Q	12	0	12	-	18
D	0	0	0	0	-

<---- PSA_3 Hierarchy ---->

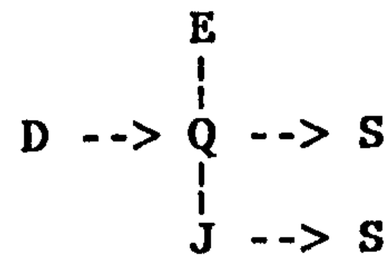


Fig 7.5.

<--- Hierarchy Matrix PSA_4 --->

	E	S	J	Q	D
E	-	0	0	0	0
S	47	-	35	24	47
J	18	6	-	18	18
Q	29	6	29	-	29
D	12	12	12	12	-

<---- PSA_4 Hierarchy ---->

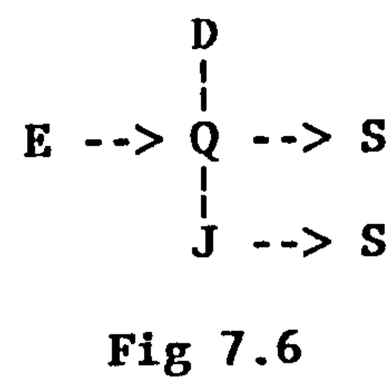


Fig 7.6

Figures 7.5 and 7.6 display the discomfirmatory hierarchy patterns for $PSA_{3\&4}$ and $PSB_{3\&4}$ (within Tables 7.22 and 7.23). A degree of similarity is evident within the structural patterns. A four profile correspondence exists between these hierarchies; profiles D and E were interchanged across the two sub-groups. Unlike the results of the inter-regression and inter-profile correlations the constructed hierarchies across PSA are apparently relatively stable in-spite of the small sample.

However, this stability is not reflected within the hierarchies associated with the $PSB_{3\&4}$ pairing (Table 7.23). Firstly, a 12% tolerance level was required for the construction of both hierarchies. Secondly, only a single profile correspondence, namely test item H, is evident across the sample. This latter result was considered to indicate a potential 'INSET effect' associated with the treatment condition X_1. Although with two groupings of only five profiles this evidence was thought to be of qualitative rather than quantative value.

Table 7.23 Disconfirmatory response patterns for two-profile prerequisite relation orderings across PSB. (N = 17)

<-- Hierarchy Matrix PSB_3 -->|<---- PSB_3 Hierarchy ---->

	H	A	T	O	K
H	-	59	65	35	53
A	0	-	12	6	0
T	0	6	-	12	0
O	6	35	47	-	35
K	12	18	24	24	-

(only possible at 12% tolerance)

A
| --> K --> O --> H
T

Fig 7.7

<-- Hierarchy Matrix PSB_4 -->|<---- PSB_4 Hierarchy ---->

	H	A	T	O	K
H	-	53	29	65	65
A	0	-	6	18	18
T	0	29	-	47	35
O	0	6	12	-	12
K	0	6	0	12	-

(only possible at 12% tolerance)

K
| --> A --> T --> H
O

Fig 7.8

7.4.3 Between Sub-Groups: multiple-regression analysis.

Results

Within both sub-groups of the York cohort, Assessment Profile variables were regressed onto individual teacher regression equations (logistic regression weights), generating 'policy to practice' correlations. These are illustrated within Tables 7.24 and 7.25. A preliminary inspection of both sub-groups across the two profile sets revealed a moderate or substantial set of combined predictor variable correlations in 7 of the 12 possible regression calculations (R^2 values). Detailed inspections were then undertaken across sub-groups and between profile-sets.

The PSA_3 regression weight data, depicted within Table 7.24, indicated a moderate correlation (R^2=0.32) with the combined predictor variables and the THRESHOLD proficiency component of the teachers' judgement policy. Similarly, a moderate correlation (R^2=0.39) was registered for the MAXIMUM proficiency component of the policy. No individual variable made a significant contribution to the prediction of either of these regression weights. Inspection of PSA_4 (Table 7.25) revealed no moderate correlations with the combined predictor variables and the components of the judgement policy.

Table 7.24 Multiple Regression of Assessment Profile Information onto teachers' regression equations (standardised coefficients). Sgp 3 (N = 17)

API	<------- PSA_3 ------->			<------- PSB_3 ------->		
	CON	THR	MAX	CON	THR	MAX
PDF	-0.32	-0.17	-0.37	-0.09	-0.06	-0.09
NDF	0.25	-0.22	0.15	0.63	-0.43	0.80
PSC	0.14	-0.23	0.20	0.01	-0.15	-0.02
WKS	0.31	-0.29	0.06	-0.20	-0.12	0.07
CNG	-0.04	0.40	0.38	0.10	-0.21	0.04
R^2	0.14	0.32	0.39	0.43	0.36	0.56

* $p<.05$ ** $p<.01$

Inspection of the PSB_3 data, illustrated within Table 7.24, revealed a moderate correlation with the combined predictor variables and the CONGRUENCE (R^2=0.43) and THRESHOLD proficiency (R^2=0.36) components of the judgement policy. Similarly, a substantial correlation (R^2=0.56) with the combined predictor variables and the MAXIMUM proficiency component was registered. No significant independent contributions were notable within any of the three combined cases. A similar consideration of the PSB_4 data (Table 7.25) indicated a substantial correlation (R^2=0.51) with the combined predictor variables and CONGRUENCE component of the judgement policy. Additionally, a moderate correlation (R^2=0.38) was also registered for the MAXIMUM proficiency component of the judgement policy. In this latter case no individual predictor variables made a significant contribution. Furthermore, the individual proficiency score variable (PSC) correlated significantly ($p<0.05$) with the CONGRUENCE component of the

judgement policy: This identified a covariation in congruence influence and a subject's perception of proficiency.

Table 7.25 Multiple Regression of Assessment Profile Sgp 4 (N = 17) Information onto teachers' regression equations (standardised coefficients).

API	<------- PSA_4 -------> CON	THR	MAX	<------- PSB_4 -------> CON	THR	MAX
PDF	0.46	-0.68	-0.00	-0.57	0.17	-0.47
NDF	0.08	0.61	0.16	0.01	-0.01	0.01
PSC	-0.14	-0.09	0.10	0.64*	0.20	-0.11
WKS	0.22	-0.31	0.17	-0.52	0.35	-0.68
CNG	-0.28	-0.11	-0.02	-0.33	0.09	-0.04
R^2	0.28	0.19	0.05	0.51	0.13	0.38

* p<.05 ** p<.01

Comparison between PSA and PSB (Tables 7.24 & 7.25) across the two sub-groups of the York cohort indicated a degree of difference between expressed policy and actual practice. Initially, this was evident through the adoption of complementary judgement policies between the two profile sets (consistent across the sample). Teachers perception of the utility of Statements of Attainment as assessment criteria (depicted through the CNG variable) and proficiency ratings (registered through the PSC variable) appeared to have minimal association with the actual influence of the CONGRUENCE status and THRESHOLD or MAXIMUM proficiency cues depicted within the student profiles.

The regression of Teacher Biographic variables onto the judgement equations (logistic regression weights), within both samples of the York cohort generated individual moderator correlations (Table 7.26 and 7.27). In common with the Assessment Profile analysis, standardised regression weights were adopted for the purposes of analysis. A preliminary inspection of both sub-groups across the two profile sets indicated a moderate or substantial set of combined predictor variable correlations in 5 of the possible 12 outcomes. Subsequently, detailed inspections were then undertaken allowing comparisons to be considered across the sample and between profile sets.

Illustrated within Table 7.26 is the PSA_3 and PSB_3 regression correlation data. The PSA_3 data indicated a moderate correlation (R^2=0.30) with the combined predictor variables and the CONGRUENCE component of judgement policy. Although, no individual variables made a significant independent contribution. A substantial correlation (R^2=0.51) was registered between the combined predictor variables and the THRESHOLD proficiency component of the policy. In this instance the National Curriculum Assessment Experience variable (NCEX) demonstrated a highly significant ($p<0.01$) correlation with the THRESHOLD proficiency component of the policy. This identified a covariation between specific NC assessment experience and proficiency influence. More specifically, greater proficiency

influence was associated with teachers' previous levels of assessment experience assessment. Inspection of PSA_3 (Table 7.27) indicated a moderate correlation (R^2=0.43) with the combined predictor variables and the THRESHOLD component of judgement policy. In this instance Teachers' Principal Subject variable (PSUB) indicated a significant ($p<0.05$) correlation with the THRESHOLD proficiency component of the policy. This identified a covariation between Mathematics as a main taught subject and THRESHOLD proficiency influence. Non-mathematicians registered a greater degree of susceptibility to the THRESHOLD proficiency cue depicted within student hypothetical profiles.

Table 7.26 Multiple Regression of Teacher Biographic Sgp 3 (N = 17) Information onto teachers' regression equations (standardised coefficients).

TBI	<------- PSA_3 -------> CON	THR	MAX	<------- PSB3 -------> CON	THR	MAX
YTEA	-0.48	-0.36	-0.19	0.20	-0.80*	0.65
PSUB	-0.06	-0.01	-0.07	-0.28	-0.34	0.19
CREX	0.10	0.18	-0.14	-0.16	0.47	-0.34
NCEX	-0.20	0.78**	0.46	-0.14	0.54*	-0.48
R^2	0.30	0.51	0.17	0.16	0.46	0.30

* $p<.05$ ** $p<.01$

When inspected the PSB_3 data, shown within Table 7.26, indicated a moderate to substantial correlation (R^2=0.43) with the combined predictor variables and the THRESHOLD proficiency component of the judgement policy. A moderate correlation (R^2=0.30) was registered for the MAXIMUM proficiency component

of the policy. Additionally, the years teaching (YTEA) variable correlated significantly ($p<0.05$), with the THRESHOLD proficiency component, greater proficiency influence was associated with the length of service. In contrast, the national curriculum assessment experience (NCEX) variable correlated significantly ($P<0.05$) with the THRESHOLD proficiency component of judgement policy. A similar inspection of PSB_4 (Table 7.27) indicated no moderate correlation with the combined predictor variables and any component of the judgement policy.

Table 7.27 Multiple Regression of Teacher Biographic Sgp 4 (N - 17) Information onto teachers' regression equations (standardised coefficients).

	<------- PSA_4 ------->			<------- PSB_4 ------->		
TBI	CON	THR	MAX	CON	THR	MAX
YTEA	-0.37	0.01	-0.54	0.13	-0.29	0.40
PSUB	-0.13	**-0.51***	0.11	0.19	-0.16	0.00
CREX	-0.08	0.36	-0.08	-0.38	-0.03	0.13
NCEX	-0.31	0.01	0.04	-0.19	-0.22	0.20
R^2	0.28	**0.43**	0.25	0.26	0.19	0.18

* $p<.05$ ** $p<.01$

Both PSA and PSB, when considered across the York sample (Tables 7.26 & 7.27), indicated a general level of inconsistency for all four potential moderator variables. However, a limited pattern of consistency was evident for the NCEX variable in particular. Experience of National Curriculum assessment (expressed through the variable NCEX) registered

consistent correlations between PSA_3 and PSB_3. Although, any meaningful interpretation of these is limited by the absence of significant cross sub-group evidence.

7.4.4 Between Subjects: homogeneity analyses.

Results

Table 7.28 depicts the regression equation frequency distribution for PSA_3; and it indicates three key judgement policies were predominantly utilised by teachers. The influence of CONGRUENCE status was prominent within two policies; which accounted for 7 judgement policies in total. Although, for one of these policies (38.2, 39.3, -1.1), three teachers also recognised the importance of THRESHOLD proficiency. Additionally, the influence of the THRESHOLD proficiency cue was dominant within the final key policy; accounting for four teacher judgements. The six remaining teachers adopted policies which were unique; each emphasising differing influences of the CONGRUENCE status and THRESHOLD or MAXIMUM proficiency cues depicted within the student hypothetical profiles. A similar inspection of the regression equation frequency distribution for PSA_4 (Table 7.28) revealed the dominance of two key judgement policies. Again the influence of CONGRUENCE status was prominent within both policies; accounting for 9 of the 17 teacher judgements in total. Although this influence was in combination with either THRESHOLD or MAXIMUM proficiency. Only

five individual judgements displayed the 'perfect solution' (38.2, 39.3, -1.1) of CONGRUENCE status and THRESHOLD proficiency cue influence. The eight remaining teachers adopted policies emphasising a combination of CONGRUENCE status and/or THRESHOLD and MAXIMUM proficiency influence.

Table 7.28 Frequency distribution of teachers' regression equations across PSA and PSB (unstandardised coefficients).

Tr	<----- PSA_3 -----> CON	THR	MAX	Fr	Tr	<----- PSB_3 -----> CON	THR	MAX	Fr
X_n	38.2	39.3	-1.1	3	X_1	38.2	39.3	-1.1	1
	38.2	1.1	37.1	1		0.0	38.4	0.0	4
	38.4	0.0	0.0	4		38.2	-1.1	39.3	2
	38.2	-37.1	-1.1	1		0.0	38.4	-38.4	1
	38.2	1.1	-39.3	1		0.0	-38.4	38.4	1
	0.0	38.4	0.0	4		0.0	-9.2	9.2	2
	0.0	0.0	0.0	1		-38.2	77.5	39.3	5
	0.0	9.2	-9.2	1		-38.2	-1.1	39.3	1
	0.0	0.0	-38.4	1					
MEAN	22.5	14.5	-3.2			-6.7	32.9	-3.6	
Tr	<----- PSA_4 -----> CON	THR	MAX	Fr	Tr	<----- PSB_4 -----> CON	THR	MAX	Fr
X_1	38.2	39.3	-1.1	5	X_n	38.2	39.3	-1.1	4
	38.4	0.0	0.0	2		0.0	9.2	-9.2	1
	38.2	-1.1	39.3	2		0.0	38.4	0.0	3
	38.2	1.1	-39.3	4		0.0	0.0	0.0	2
	38.2	-37.1	-1.1	2		0.0	-38.4	38.4	1
	0.0	38.4	-38.4	1		-38.2	39.3	-1.1	4
	0.0	9.2	-9.2	1		-38.2	77.5	-39.3	1
						-38.2	-1.1	39.3	1
MEAN	33.7	10.1	-7.9			-4.5	28.0	1.2	

(bold print indicates judgement policy frequency > 2)

A consideration of the regression equation frequency distribution for PSB_3, shown within Table 7.28, revealed two

key judgement policies were utilised by teachers. The influence of the THRESHOLD proficiency cue was prominent in both policies; involving 9 of the 17 teacher judgements in total. Although, for one of these policies (-38.2, 77.5, 39.3), 5 teachers also realised the importance of CONGRUENCE status with THRESHOLD proficiency. The eight remaining teachers adopted policies which emphasised differing influences of the judgement components. A similar examination of PSB_4 (Table 7.28) revealed the utilisation by teachers of three key judgement policies. The influence of the THRESHOLD proficiency cue was prominent in all three policies; accounting for 11 of the 17 teacher judgements in total. Although for two policies (-38.2, 39.3, -1.1) and (38.2, 39.3, -1.1) 8 teachers also depicted this THRESHOLD proficiency influence in combination with that of CONGRUENCE status. Only four teachers were identified with 'perfect solution' (38.2, 39.3, -1.1) decision strategies. The six remaining teachers adopted policies emphasising differing levels of influence for CONGRUENCE status and THRESHOLD/MAXIMUM proficiency.

Comparison between PSA and PSB across the two sub-groups of the York cohort indicated the judgement policies utilised tended to be complementary. Across both profile sets the prominent decision strategies adopted by teachers appeared to demonstrate a mutually exclusive use of the dimensional information within student hypothetical profiles. Regarding PSA teachers, decision

strategies were dominated by CONGRUENCE status; although influence of the THRESHOLD and MAXIMUM proficiency cues was also detectable, if somewhat variable. In contrast, PSB revealed teachers' appeared to utilise decision strategies with a prominent THRESHOLD proficiency cue influence; although a variability of CONGRUENCE status influence was also discernible. Of the 64 policies illustrated 13 depicted the 'perfect solution' of a balance judgement (38.2 39.3 -1.1). The between-subjects homogeneity analysis confirmed the acceptability of reporting an overall regression equation for sub-group 3; but not 4. Finally, the significant differences of homogeneity within $PSA_{3\&4}$ and $PSB_{3\&4}$ is interpretable as a potential 'In-Set effect' associated with the X_1 treatment condition.

7.4.5 Discussion of the findings.

The exploration of 'teacher professional judgement' in the context of an In-Service Education and Training treatment condition revealed evidence for a modified schema-based cognitive simplification strategy. The results of the logistic regression analysis demonstrated a post-treatment enhancement of CONGRUENCE status and THRESHOLD proficiency influence within the complementary judgement policies adopted within PSA and PSB. Similarly, a discernible increase in the negative influence of the MAXIMUM proficiency cue was detectable within

this post-treatment (INSET) condition. However, the presence of a potential 'INSET effect' did not appear to compromise the stability within profile-sets and across the sample of the inter-regression weights. Although, a potential 'INSET effect' was a noticeable feature of the ordering theoretic hierarchical analysis results.

The nature of the judgement policies, utilised by teachers, was further delineated by the multiple-regression analysis. Specifically, implicit policy and explicit practice differences were evident. In addition, this took the form of a redundancy of association between policy and practice, rather than that of a discernible inappropriate relationship. The potential utility of teacher biographic information to provide an explanation for decision strategy differences was indicated within the post-treatment conditions. However, in the absence of a consistent pattern of significant independent contributions any practical benefits were restricted to combinations of predictor variables. Finally, the variation of judgement policies between the two profile sets revealed the adoption of an homogenous decision strategies within sub-group 3 (across PSA and PSB) but not sub-group 4. This latter result providing evidence of a potential pre-test/post-test 'INSET effect'.

7.5 Modification of a cognitive simplification strategy.

The results of the pilot study analysis highlighted the difficulty associated with the determination of any significant 'treatment effect'. Consequently, the main study analysis, involving several treatment conditions and small sample sizes was revised and undertaken within a modified format. This provided the opportunity for analyses, supportive of the pilot study, to be considered. As a consequence, it was also possible to achieve a reduction in the limitations associated with the Quasi-Experimental Research Design adopted for the main study.

The analysis modifications involved three aspects. The first concerned the small sub-group sizes. To compensate for this, 'pooling' of sub-groups 5 with 6 and 7 with 8 was undertaken for certain analyses. This increased sample sizes, both combined sub-groups became 24 in number. Secondly, it was anticipated that the measurable difference between treatment conditions X_2 and X_3 (both were variations of X_0) would of a negligible size; hence both could be reduced to X_0. Similarly, the measurable difference between X_4 and X_5 (both were variations of X_1) was expected to be negligible, therefore a reduction to X_1 was possible. Thirdly, the range of analyses was reviewed. The logistic regression element was retained and incorporated within a cluster analysis. The correlation and hierarchical analyses were removed. The multiple regression analysis, involving teacher biographic information was

undertaken and is summarised within Chapter 8. Similarly, the homogeneity analysis was completed and is also summarised within chapter 8.

The samples identified for the fulfilment of this supportive analysis were sub-groups 5, 6, 7 and 8. The testing arrangement utilised a variation on a quasi-experimental non-equivalent control group design:

O_{5a} X_0 O_{5b}

O_{6a} X_0 O_{6b}

O_{7a} X_1 O_{7b}

O_{8a} X_1 O_{8b}

non-pooled version

$O_{5\&6a}$ X_0 $O_{5\&6b}$

$O_{7\&8a}$ X_1 $O_{7\&8b}$

pooled version

The depiction O_{5a}, for example, represents the observed ratings for PSA, given by the 5th sub-group. The condition X_1, for instance, designates the INSET treatment condition.

7.5.1 Within Sub-groups: cluster analysis.

Procedure

The cluster analysis technique utilised the teacher as the unit of analysis. The collection of regression equations (112 across the eight sub-groups) were grouped into a pre-designated number of clusters. Cluster centres were generated which minimised the distance between each regression equation and its nearest

centre. Within this statistical procedure, the determination of cluster significance is a matter of judgement. Although several methods of judgement are available, the extent to which group membership (of each cluster) discriminates against an external variable was considered to be the most appropriate. The association of clusters with sub-group designation was then explored.

Table 7.29 Cluster membership and centres across PSA and PSB for the pilot and main study teacher regression equations (unstandardised).

	<--------- PSA --------->									<-Cluster Centres->		
Cls	X_n 1	X_o 2	X_n 3	X_1 4	X_n 5	X_n 6	X_n 7	X_n 8	Sum	CON	THR	MAX
1	1	2	3	1	4	1	2	0	14	0.0	1.9	-7.4
2	8	2	1	2	2	0	1	1	17	38.2	-1.5	37.3
3	0	1	3	5	0	1	1	3	14	38.2	39.0	-1.0
4	0	1	0	0	0	0	0	0	1	-38.2	39.0	-1.0
5	0	1	1	4	0	1	1	0	8	38.2	1.0	-39.0
6	1	0	0	0	0	1	0	0	2	-38.2	77.0	-39.0
7	1	0	4	1	0	1	1	0	8	0.0	38.0	-9.5
8	4	8	5	4	5	8	7	7	48	38.3	-10.8	-0.3

	<--------- PSB --------->									<-Cluster Centres->		
Cls	X_o 1	X_n 2	X_1 3	X_n 4	X_o 5	X_o 6	X_1 7	X_1 8	Sum	CON	THR	MAX
1	0	0	4	2	1	0	2	0	9	-8.4	-19.1	31.8
2	1	0	5	1	0	2	1	0	10	-38.0	77.0	-39.0
3	0	0	2	0	1	0	1	0	4	38.0	0.0	38.0
4	13	11	5	6	4	8	8	6	61	0.0	34.1	-2.9
5	0	3	0	4	4	3	0	2	16	-38.0	38.9	-0.9
6	0	0	0	0	0	0	0	2	2	38.0	-37.0	-1.0
7	0	1	1	4	1	0	0	0	7	38.0	39.0	-1.0
8	1	0	0	0	0	0	0	1	3	-38.0	25.7	-38.3

Tot'	15	15	17	17	11	13	13	11	112

(Cls - Cluster number)

Results

An inspection of the cluster membership data for PSA represented within Table 7.29, indicated the dominance of one judgement policy (38.3, -10.8, -0.3). The influence of CONGRUENCE status was prominent within this decision strategy. Prominence was defined as a regression value of approximately 30 (or greater), the size associated with exact fit equations within the logistic regression analysis. The importance of the CONGRUENCE cue was evident within 90 of the 112 designated judgment policies. Any differences between the two treatment conditions X_o and X_1 with X_n were not apparent from a visual inspection of the cluster membership data (Table 7.29). No consistent evidence was found across sub-groups for the determination of either a 'learning-effect' (X_o treatment condition) or an 'INSET effect' (X_1 treatment condition). A consideration of the PSB data revealed the dominance of one judgement policy (0.0, 34.1, -2.9). The influence of the THRESHOLD proficiency cue was prominent within this decision strategy. The importance of this cue was apparent within 96 of the 112 judgment policies illustrated. Finally, in common with the PSA findings, no consistent differences between the two conditions X_o and X_1 with that of X_n (non-treatment) were evident from a visual inspection of the cluster membership data (Table 7.29).

Comparison between PSA and PSB across the pilot and main study data indicated the judgement policies utilised tended to be complementary. Across both profile sets the prominent decision strategies adopted by teachers demonstrated a mutually exclusive use of the dimensional information within student profiles. For PSA, teachers decision strategies were dominated by CONGRUENCE status; influence of the THRESHOLD and MAXIMUM proficiency cues was evident though, if not variable. In contrast, PSB revealed teachers' appeared to utilise decision strategies with a prominent THRESHOLD proficiency influence; although a variability of CONGRUENCE status and MAXIMUM proficiency influence was also apparent. Of the 224 policies illustrated 21 depicted the 'perfect solution' of a balanced judgement (38.0, 39.0, -1.0).

7.5.2 Discussion of the findings.

Although the cluster membership analysis was based on subjective judgement, it did allow the preliminary uni-dimensional analysis findings (section 7.1) to be investigated within the context of a multi-dimensional diagnostic procedure. The results of the t-test analysis (Table 7.5) indicated differences within sub-groups 4 and 8 across PSA and PSB. Consideration of the respective cluster membership data (Table 7.29) illustrates the nature of these differences. For instance, sub-group 4 depicts a definitive contrast between the

adopted judgement polices between PSA and PSB. CONGRUENCE status is prominent in 14 of the judgement policies given for PSA. However, this is important for only 9 teachers within PSB. A similar contrast is evident within sub-group 8. CONGRUENCE status is dominant within all 11 judgements for PSA. However, within PSB this importance is reflected in only 5 decision strategies.

The results of the ONEWAY analysis (Table 7.6) indicated significant differences between sub-groups 4 & 5 and 5 & 8 (for PSA). These differences are evident within the cluster membership data (Table 7.41). Sub-groups 4, for example, indicates 5 teachers utilised the 'perfect solution' judgement policy. This is not illustrated within sub-group 5. Similarly, sub-group 5 revealed only 7 (out of 11) teachers were influenced by the CONGRUENCE status. In contrast, 10 teachers were influenced by this cue within sub-group 8.

In general, the clustering technique applied to the data for the pilot and main studies provided an effective means of diagnostic analysis. The significant differences illustrated by the preliminary data analysis, although not confirmed by this latter analysis, were detailed in terms of their underlying decision strategies. Hence, the cluster analysis fulfilled its

principal aim of describing the nature of the (significant) differences in terms of individual teacher judgment policy.

7.5.3 Summary.

The preliminary analysis of the data collected within the pilot and main studies indicated significant differences between subjects in terms of their criterion-referencing ability (or skill). These differences were investigated through a series of aims and associated hypotheses. The intention was to determine the mechanism by which teachers made decisions. The analysis of judgement policies within the pilot study data provided a range of diagnostic measures and indicated possible explanations for the apparent differences. Problems of sample size, treatment condition differentiation and experimental design considerations compelled the main study analysis to adopt a revised role. The modified analysis of the main study data provided additional, supportive evidence for the findings obtained from the earlier pilot study analysis. The final chapter will summarise and discuss both the results and findings obtained within the pilot and main study analyses, concluding with a consideration of the 'policy-capturing' model as a future research design.

Chapter 8.

A Cognitive Model of the Judgement Process: evaluation and conclusions of the study.

The aims of this concluding chapter are three fold. Firstly, it will provide a summary of the problem, methodology, and results. Secondly, it will describe and interpret the findings in the context of previous research; highlighting the limitations of the research design and methodology. Finally, the general findings will be discussed and summarised; with suggestions for future research strategies provided.

8.1 Introduction.

The use of Criterion Referenced assessment, within the National Curriculum, brought into focus the concept of 'Teacher Professional Judgement'. The judgement process involves two complementary aspects. The decision-making strategy, or rating-policy, with which judgements are undertaken by the teacher is the first. The second concerns the Statements of Attainment, or rating-criteria, against which the judgements are referred. The interaction between the teacher's judgement policy and the commensurate assessment criterion on which this is applied, formed the basis of the researchable problem within the context of this study.

The aims of the study addressed three distinct aspects. Firstly, the formulation of a cognitive-model of the judgement process using the findings of previous research was considered. Secondly, the modification of this cognitive-model through In-Service Training was undertaken. These two aims formed the basis of the Pilot Study. The final third aim was complementary to the first two, it introduced assessment environments as an additional variable to be investigated. This aim was to have been the basis of the main study. Initially, six hypotheses were formulated to enable the fulfilment of the research aims. Small sample sizes within the main study required the redefinition of purpose for this aspect. Consequently, this aim became a supportive element of the first two aims with a corresponding reduction in hypotheses to four.

The methodology was based on a 'policy-capturing' design (variation), utilised previously within research on teacher judgements. The data collection had four key aspects. Two questionnaires collected data of a factual and opinion based nature. The third aspect involved two sets of five student profiles, each depicting three information cues. A series of Statements of Attainment to which the student profiles were referenced and the dichotomous ratings recorded, formed the fourth and final aspect. Within the pilot study, the test-battery was utilised within two North Yorkshire venues as part of an In-Service Training session. Within the main study the

test-instrument was distributed to a sample of Humberside secondary schools for completion and postal return.

8.2 The Findings: a summary of the results.

The main findings of the results of the investigation are shown within Figures 8.1 to 8.5. These relate to the two aims investigated and four hypotheses tested. The teacher judgment analyses are illustrated across the pilot and main studies and between the treatment and non-treatment conditions.

8.2.1 Cognitive simplification strategies.

The common decision strategies identified within the teacher response patterns are illustrated within Figure 8.1. These are represented within three distinct formats. Each common response pattern is illustrated with its associated regression equation. The common decisions are reduced to congruence and proficiency component patterns, depicting the contribution of each separately. The theorised mechanism for the combination (or superposition) of the congruence and proficiency component patterns is depicted. Common decision strategies are defined arbitrarily as those utilised by more than 20% of a specified sample. For sub-groups 1, 2, 3 and 4 (sample sizes of 15 to 17) this became three or more teachers. For the larger combined

samples of 24 (sub-groups 5 & 6, 7 & 8) this figure was five or more teachers.

Fig. 8.1 Common decision strategies (>2) within each pilot study sub-group, (>4) within each main study combined sample.

Regression equation & response pattern	Congruence & Proficiency	Superposition (C_x * P_y)	N=
38.2 1.1 37.1 (1 1 0 1 1)	C_{cr} 01011 M_{cr} 10010	OR -> 11011	8
38.4 0.0 0.0 (0 1 0 1 1)	C_{cr} 01011 P_o Identity	ID -> 01011	32
38.2 39.3 -1.1 (0 0 0 1 1)	C_{cr} 01011 T_{cr} 10011	AND -> 00011	8
0.0 38.4 0.0 (1 0 0 1 1)	C_o Identity T_{cr} 10011	ID -> 10011	54
38.2 1.1 -39.3 (0 1 0 0 1)	C_{cr} 01011 M_{ir} 01101	AND -> 01001	4
38.2 -37.1 -1.1 (0 1 1 1 1)	C_{cr} 01011 T_{ir} 01100	OR -> 01111	6
-38.2 39.3 -1.1 (1 0 1 1 1)	C_{ir} 10100 T_{cr} 10011	OR -> 10111	13
-38.2 77.5 -39.5 (1 0 0 0 1)	C_{ir} 10100 P?	Non-interpretable	5
		Total = 130/224 (2x112)	

(* represents the operations of AND, OR and ID)

The policy capturing technique provided two distinct elements of the teacher decision strategy. The first element related to 'what' informational cues teachers considered to be of importance. The congruence cue was evident within two differentiated levels:

(i) Congruence recognition, which was sub-divided into correct and incorrect judgment, denoted by C_{cr} and C_{ir} respectively;

(ii) congruence non-recognition which was denoted by the identity reference C_o;

The proficiency cue was apparent within three differentiated levels:

(i) Threshold proficiency recognition, which was sub-divided into correct and incorrect judgement, denoted by T_{cr} and T_{ir} respectively;

(ii) Maximum proficiency recognition, which was sub-divided into correct and incorrect judgement, denoted by M_{cr} and M_{ir} respectively;

(iii) Proficiency non-recognition which was denoted by the identity reference P_o;

From the descriptions of each cue level of influence it is possible to characterise the judgement process. For example, the cue designation C_{ir} represents a teacher with the skill of congruence recognition, but this awareness is allied to an incorrect judgement determination (ie. assigns negative judgments to profiles with congruence and positive judgments to profiles with non-congruence). Similarly, a cue designation M_{cr} represents a teacher with the skill of maximum proficiency recognition and this awareness is allied to the correct judgement determination (ie. assigns positive judgements to profiles with maximum proficiency and negative judgements to

profiles with non-maximum proficiency). An identity cue designation, for instance P_o, represents a teacher without the skill of proficiency recognition (ie. positive and negative judgements are not associated with proficiency cue status).

The second element of the judgement policy illustrated 'how' the cues were combined. An 'effective theory' was developed which involved the superposition of the individual congruence and proficiency component patterns. Although the combination of the individual cue influences may not 'in actuality' occur in terms of the superposition of distinct congruence and proficiency patterns the outcome was consistent with the results. Three distinct superposition categories were evident within the decision strategy patterns. The first was compensatory in nature. The judgment decision reflected the influence of either the congruence or proficiency cue but not necessarily both. Mathematically this is equivalent to a 'logical-OR' (OR) operation. The second was exhaustive in form. Both congruence and proficiency cues were essential and necessarily influential within the judgement decision. Mathematically, this is equivalent to a 'logical-AND' (AND) operation. The third category related to an 'identity' (ID) combination. Either the congruence cue was dominant with proficiency in a redundant (identity) state or the reverse occurred. This evidence of interactive cue influence is consistent with the findings reported by Slovic and

Lichtenstein (1971) relating to the difficulties judges encounter weighting and combining information.

The response pattern information illustrated within Figure 8.1 indicate 125 out of 130 common decision strategies were associated with the combinations of 'AND', 'OR' and 'ID' which provides a high degree of credibility and validity for the 'effective theory' of congruence and proficiency pattern superposition. The two prominent identity decision strategies illustrated (C_{cr} * P_o and C_o * T_{cr}) account for 86/130 of the common judgements. The secondary decision strategies (relating to the AND/OR combinations) although individually modest in frequency still account overall for 39 judgements. In contrast, only 5 judgements were associated with a non-interpretable combination of congruence and proficiency cue influence. The small percentage of non-interpretable deviations from the effective theory provides further support for the utility of the superposition concept (of congruence and proficiency component patterns) developed within this study.

8.2.2 Heuristic decision strategies.

Within figure 8.2 are the findings associated with the Assessment Profile Information multiple regressions. The results generally indicate that congruence and proficiency influences implicit in the ratings given by teachers were not

evident within the explicit policies recorded. This is demonstrated by the absence of the predictor variables of CNG and PSC for the influence of the congruence and proficiency cues respectively.

Fig.8.2 Assessment Profile Information Summary: Combined predictor and individual variable influence for all sub-groups.

Student Cue	sub-group 1	sub-group 2	sub-group 3	sub-group 4
PSA	**X_n**	**X_0**	**X_n**	**X_1**
CON	34%	88% PSC*		
THR	52%	76% -WKS*		
MAX		36%		
PSB	**X_0**	**X_n**	**X_1**	**X_n**
CON		59% WKS*	43%	51% PSC*
THR		59% -WKS*	36%	
MAX		59% WKS*	56%	38%

The range of the potential heuristic strategies adopted by teachers is shown through the presence of unexpected predictor variable covariations. For instance, the PSC rating significantly predicted the influence of congruence. Similarly, teacher perceptions regarding short-term-retention represented by the WKS variable significantly predicted congruence and

proficiency cue influences. These results indicate that teachers decision strategies may be influenced by internalised pre-conceptions. This sort of effect has been demonstrated within studies of other types of teacher judgements, for example Pedualla, Airasian and Madaues (1980).

8.2.3 Biographic moderator variables.

Within figure 8.3 are the findings associated with the Teacher Biographic Information multiple regressions. The results indicate that teachers' background characteristics were predictive of both congruence and proficiency influences within decision strategies. The extent of the predictability of the teacher background characteristics is evident through the presence of several predictor variable covariations. For example, the YTEA rating significantly predicted the influence of the threshold proficiency cue within one In-service treatment group. The negative covariation may be interpreted as indicating a greater length of service reduces the effectiveness of the applied INSET treatment. The significant appearance of the PSUB variable provides a consistent pattern of prediction. Non-mathematicians appeared to be more susceptible to the influence of the MAXIMUM proficiency cue in the absence of INSET. However, within INSET sub-groups non-mathematicians demonstrated greater susceptibility to decision strategy modification. Conversely, mathematicians were less

susceptible to decision strategy modification through INSET treatments. Assessment experience appeared to have a beneficial effect on the adopted teacher decision strategies. These effects were evident in two specific ways. The greater experience of National Curriculum assessment appeared to have a positive effect on the influence of THRESHOLD proficiency. In contrast, a lesser experience of Criterion-Referenced assessment was associated with more susceptibility to the influence of the MAXIMUM proficiency cue.

Fig.8.3 Teacher Biographic Information Summary: Combined predictor and individual variable influence for all subgroups.

Student Cue	sb-gp 1	sb-gp 2	sb-gp 3	sb-gp 4	sb-gp 5&6	sb-gp 7&8
PSA		X_n	X_o	X_n	X_1	X_n	X_n
CON		36%	30%			
THR			51% NCEX**	43% -PSUB*		
MAX	34% -PSUB*	66%			34% -CREX*	
PSB	X_o	X_n	X_1	X_n	X_o+	X_1+
CON		33%				30%
THR		33%	46% -YTEA* NCEX*			
MAX	33%	30%			30% PSUB*	

(X_o+ and X_1+ are reduced treatment conditions: see p202)

The results overall provided a pattern consistent with the expectation of a proficiency based schema underlying the decision strategies of teachers. Additionally, covariation of individual predictor variables with the CONGRUENCE status cue was not evident. Teacher decision strategies have been shown to be susceptible to modification to INSET. However, more importantly, the effectiveness of the INSET was found to be dependant on certain teacher background characteristics. This latter finding is in contrast with those of Borko and Cadwell (1982), who concluded that 'global' teacher characteristics were unrelated to teacher decision strategy differences.

8.2.4 Decision strategy homogeneity.

The homogeneity analysis (of common slopes) results, depicted within figure 8.4, demonstrate no discernible pattern for either treatment or non-treatment conditions. The expectation that an INSET effect would be evident through the increased homogeneity of rating policies is not confirmed. However, the degree of homogeneity indicated within certain sub-groups in the absence of INSET is consistent with previous research findings. For example, Greenen and Smith (1981), Graham (1989), have reported on the reliability of teacher assessments or ratings. Nevertheless, other previous research (Hoge and Butcher, 1984) has indicated a cautious approach is required with the pooling of judgemental data across teachers. Hence,

the homogeneity analysis results need to be viewed within the context of additional, supportive findings.

Figure 8.4 Decision strategy homogeneity depicted through an analysis of common slopes (F-test) for all sub-groups.

Student Cue	sb-gp 1	sb-gp 2	sb-gp 3	sb-gp 4	sb-gp 5&6	sb-gp 7&8
PSA	X_n	X_o	X_n	X_1	X_n	X_n
Status	n-s	n-s	n-s	p<.05	p<.05	p<.05
PSB	X_o	X_n	X_1	X_n	X_o^+	X_1^+
Status	n-s	p<.05	n-s	p<.05	n-s	n-s

(X_o^+ and X_1^+ are reduced treatment conditions: see p202)

8.2.5 INSET and decision strategies.

The results depicted within figure 8.5 indicate the decision strategies adopted by teachers tend to have primary and secondary features, reflecting the relative importance of congruence and proficiency. Within PSA, the principal rating policy of $C_{cr}*P_o$ indicated the congruence cue is dominant, whereas for PSB, the predominant cue for the principal rating policy of C_o*T_{cr} is that of proficiency. These primary components of the respective judgement policies appear to be very stable.

Fig. 8.5 Teacher Judgement Policies (>2) depicted within each pilot study sub-group, (>4) depicted within each main study combined sample.

Decision Policy	Logical-AND/OR	s-g 1	s-g 2	s-g 3	s-g 4	s-g 5&6	s-g 7&8
PSA		X_n	X_o	X_n	X_1	X_n	X_n
$C_{cr}*M_{1r}$	L-AND				4		
C_o*T_{cr}	ID			4			
$C_{cr}*T_{cr}$	L-AND			3	5		
$C_{cr}*T_{1r}$	L-OR					6	
$C_{cr}*M_{cr}$	L-OR	8					
$C_{cr}*P_o$	ID	3	7	4	7	11	
PSB		X_o	X_n	X_1	X_n	X_o^+	X_1^+
$C_{1r}*P_{..}$	Nil			5			
$C_{1r}*T_{cr}$	L-OR		3		4	6	
C_o*T_{cr}	ID	13	11	4	3	7	12
Sub-group Totals		15	15	17	17	24	24

(X_o^+ and X_1^+ are reduced treatment conditions: see p202)

The In-Service treatment has no demonstrable effect on these cue influences. For example, across sub-groups 1 to 4 within PSA, $C_{cr}*P_o$ shows no substantial differences between treatment conditions. This finding was both expected and hoped for; these dominant cue influences represented accurate designations of the respective components C_{cr} and T_{cr} albeit exclusively. In contrast, the secondary cue influences (for instance, $C_{cr}*M_{1r}$ or $C_{1r}*T_{cr}$) were anticipated to be susceptible to In-service

effects. However, such findings are not entirely apparent on first inspection of the results within figure 8.5, consider $C_{cr}*T_{cr}$ (PSA) across sub-groups 3 and 4 or $C_{tr}*T_{cr}$ (PSB) across sub-groups 2 and 4. After closer inspection minimal secondary In-Service influence maybe discerned. Comparison of sub-groups 1, 2, 3 with 4 suggests an In-Service increased utilisation of the logical-AND strategies $C_{cr}*M_{tr}$ and $C_{cr}*T_{cr}$ apparent across PSA.

The effectiveness of In-Service related materials, as adopted within this study, are not without precedent. Jasman (1987), reported the difficulties of In-Service programmes utilised within the development of teacher assessment skills. Specifically, deficiencies with the processes and outcomes associated with such training were found. The In-Service materials, within this study, may themselves be insufficiently sophisticated. They concentrate on 'symptoms' and not 'causes' and therefore may not address the underlying dimensions of the decision-making processes adopted by teachers. A focus on general assessment strategies and techniques (McGuiness, 1987) is one possible approach and solution.

8.3 Limitations: the methodology and results.

The chief limitation centres upon the sampling techniques utilised within the pilot and main studies. The sample sizes involved were comparable with those utilised within previous

policy capturing research (p44, for example). However, they were consistently towards the lower end of the reviewed range of sample sizes. The use of 'intact' groups within the North Yorkshire cohorts was partly compensated by the randomised aspect of treatment group allocation. Similarly, the small sample sizes of the four Humberside cohorts was overcome again to some extent by their combination into two samples. Consequently, inspection of sample characteristics across a range of biographic variables revealed the samples could be considered to be representive of the teaching population in general. This indicated a reduction in the sampling limitations, but not evidence of a total accommodation or elimination of these.

The test-instrument was a potential source of several areas of concern. Firstly, the instrument was complex. The inclusion of up to four distinct sections could have been a source of confusion; leading to misrepresentative ratings. The non-equivalence of the two profile sets (PSA & PSB) was another issue of consideration. In-spite of the extensive validity and selection process involved in the production of parallel sets of profiles the distinct response pattern results of PSA and PSB indicate the contrary. Possibly, the different subject content domains adopted for each profile within a set may have been responsible for the non-equivalence. Despite the complexity of the instrument and the profile set differences,

however, only one error (across 112 subjects) was detected within those sections where such a mistake would be evident.

The collection of data posed further problems. The amount of time allocated to the judgement process may have been inadequate for the purpose of accurate ratings. On the other hand, the adoption of an extended period of time would have caused a threat to external validity (i.e. the process should reflect a natural assessment scenario within the classroom where time is limited). The Main Study suffered from a low postal response return. This was with the use of a follow-up procedure. This highlighted the final cause of concern, teacher motivation and commitment. The allotted period for the collection of data was during a time within which the National Curriculum was undergoing extensive revisions. Mathematics teachers were addressing issues associated with a curriculum initiative undergoing substantial re-writing (SEAC, 1991).

The final limitations associated with this study are concerned with the analysis of the collected data. The use of only five test-items (each designated as a student profile) with dichotomous response, provided a limited range of rating data per teacher. With the occurrence of distinctly different response patterns, and therefore judgment policies, between PSA and PSB, the generalisability of the findings to a broader domain of student profiles is open to question, although, the

reliability findings indicate the adopted policies were stable at an individual teacher level. Both predominant and distinct policies were consistent within sub-groups across profile sets. The stability and consistency of adopted judgement policies, together with the significant differences between the two profile sets indicates a potential short-coming of the cognitive-model used within this study. However, an increase in test-items within each profile set could create a cancelling effect thereby averaging out these evident and real differences of test-item response patterns. Perhaps the use of only three cues is inadequate to capture the raters' judgment policy.

8.4 Discussion.

Possibly the most important finding within this study was the fact that teacher decision strategies could be characterised in terms of predetermined informational cues. The identification of specific decision strategies which feature congruence and proficiency cue influence has implications for the conceptualization of judgements within teacher assessment. The range and stability of decision strategies indicated within the regression analysis section of this study suggests that teacher decision strategies function with a degree of complexity hitherto not considered.

The finding of significant proficiency cue influence supports the original hypothesis that decision strategies are a schema-based process constrained by the teacher's information-processing abilities. The significant congruence influence recorded within certain aspects of the regression analysis identifies a discrepancy within the cognitive-model derived for use within this study. Possibly the model is incomplete or inadequate in its ability to describe accurately the judgment process in the context of teacher assessment. The incorporation of additional informational cues may be needed to provide a more complete description of the judgement process. The finding of two different and mutually exclusive judgement policies evident between the separate profile sets used within the study may represent two facets in need of unification to form a more comprehensive model of teacher decision strategies.

The effects of INSET designed to modify teacher decision strategies were found to be evident within several aspects of the analysis. Although these effects were not always helpful in terms of resultant assessment practices, they did provide evidence for the utility of in-service training. The modification of individual cue influence enables training to be specifically directed at the cause of the problem rather than at the symptoms. Essentially, the INSET package piloted within this study formed the basis of an awareness raising exercise. This approach has been shown to have an effect illustrated by

the findings associated with the visual inspection of the derived regression equations (for individual teachers). However, the evident modification to the influence of individual congruence and proficiency cues was found to be difficult to quantify by means of statistical analysis. Individual teacher background variables were considered likely to be responsible for confounding effects. The issue of background variables illustrates the problems associated with results derived from experimental research. Natural settings, for instance within classrooms, compel the researcher to account for additional variables which may provide confounding factors or biasing. For example, variables associated more directly with student assessment within the classroom and any related qualities, experiences or skills of the teacher involved. The possible importance of these variables reinforces the need to incorporate additional cues within the cognitive-model developed through this study if the transfer from experimental to a more natural setting is to be successful.

Although the identification of individual cue influence provides important information for the characterization of a teacher's decision strategy, it is incomplete without consideration of how these cues are integrated to provide a judgment. Several options are possible for this process, for instance the integration could occur in a sequential manner, or be compensatory, or all cues could be simultaneously

assimilated to provide a judgement. The findings obtained from this study identify the integration process to be predominantly compensatory but with evidence of an exhaustive element indicated within certain decision strategies. In particular, the dominant compensatory decision strategies highlighted the schema-based covariation involving only one of the congruence or proficiency cues. In contrast, exhaustive assimilation required both congruence and proficiency cues to be evident within the judgement process.

The homogeneity of judgement policies across a range of teachers illustrated an area of concern regarding the aggregation of individual regression equations. The representativeness of a judgement policy derived from the pooling of individual teacher decisions was found to be problematic. Incorrect pooling may distort the combination process and lead to cancelling of specific elements of individual decision strategies. In this respect the utilization of 'policy-capturing' research methods requires the careful consideration of the means by which collected data are analysed. Similarly, the differences evident from the visual inspection of the regression data require careful analysis if these are to be detectable to the degree of statistical significance using for instance a clustering technique.

The development of an 'effective theory' for the decision strategies of teachers indicated one area of progress made by this research. The use of compensatory and exhaustive integration of student cue information, whilst not comprehensive in its characterisation of teacher decision strategies, marks a 'baseline' for the future development of a more effective cognitive-model of the judgement process. A remarkable feature of this 'effective-theory' approach is the degree to which it has been shown to be productive within this research. Although, it was expected that the use of the INSET package would have been more effective in the promotion of the exhaustive approach to teacher integration of student cues. The effectiveness of the package may have been compromised by the background variable differences not included in decision strategy regression equations. The possible effects of unknown background variables highlights the difficulties associated with dealing with decision strategies within the natural classroom setting. The interactive effects of students and teachers together with environmental or resource factors may produce confounding of the cue integration process within any decision strategy.

The reasons why teachers make the decisions they do was found to be varied. Importantly, this research indicated that certain aspects of the judgement process were undertaken intuitively. These heuristic decision strategies are capable of either

enhancing the judgement process or detracting from it. Unfortunately, the covert nature of such strategies make their effects unpredictable, although their detection was possible. The fact that expressed judgement policies did not predict actual practice was expected. However, the use of a series of variables which correlated with teachers' actual practice provided evidence for the possible existence of an extended schema-based judgment process. An unexpected feature of the heuristic strategies underlying teachers' judgments was the importance of the short-term-retention and proficiency score variables. These variables may be thought of as 'benchmarks' for the influence congruence and proficiency cues have within the judgement process. Clearly, these 'benchmark' variables have a significant place within the schema-based judgment process. However, the extent of this importance and the specific function(s) of these variables remain unanswered within this study.

Finally, the least expected finding within this study related to the effect of teacher biographic variable differences. Although these background differences were considered sufficiently important to warrant investigation, their impact on the decision strategies of teachers was underestimated. It was expected that the effects of INSET, for instance, might be obscured by teacher background variable differences. However,

it was not anticipated that background variables would be predictive of specific aspects of the judgment process.

The fact teaching experience had both a positive and negative influence on decision strategies and that these two opposing effects could be identified was a significant outcome of this research. The negative effect of general teaching experience was evident in the resistance exhibited to modification of decision strategies. The stability of such judgement policies precluded the important process of evaluation, review and revision and simply perpetuated any 'bad practice'. In contrast, the positive aspect of specific experience, relating to assessment within the national curriculum for instance, was apparent in the effectiveness of the associated teacher decision strategies.

8.4.1 Implications.

Professional judgement forms an important part of education at all levels from the assessment of student work undertaken within the confines of the classroom to the appraisal of a member of staff during the delivery of a lesson. The existence of judgement processes which are schema-based has been hypothesized and investigated for many years. However, the possibility of identifying the underlying schematic effects and

compensating for these is a departure from the findings of the contemporary research.

The ability of the cognitive-model developed within this research to characterise teacher decision strategies has initial implications for pre-service teacher training. Similarly, the evidence of heuristic strategies operating within the judgement process of established teachers has ramifications for the mentoring made available to newly qualified and trainee teachers. If mentoring is to be a worthwhile pursuit then the problems associated with heuristics have to be addressed and alleviated. Teacher biographic variables have several implications for the professional, in-service development of assessment practices within schools. The effective provision of INSET requires individual teacher profiling to be undertaken and the implementation of targeted training which will address the causes of deficiencies within decision strategies and not simply the symptoms.

Finally, the move from experimental research findings to natural classroom based practice will require further investigation. The development of comprehensive models which incorporate a more extensive range of classroom related variables is one consequence of any future transitional research. The possibility of teachers or educators having knowledge of their own decision strategies and the ability to

modify these to promote effective assessment practice not only offers a means to enhance professionalism but also increases the credibility of professional judgement as an accepted practice. In this respect it is in the interest of government, local authorities and examination consortia to invest in professional judgement both in terms of further research and pre- and in-service provision at all levels of education.

8.4.2 Summary.

The aim of understanding the 'what'; 'how'; and 'why' aspects of the judgement process undertaken within this study has been achieved. A model of the judgement process has been developed and indicates a cognitive-simplification strategy to be in operation. This cognitive-model, together with the tentative explanation for the differences which exist between teachers rating policies could be instrumental in the promotion of greater professional standards within assessment, education or otherwise. The ability of teachers to confront their own limitations combined with the availability of appropriate in-service training is a powerful combination. The findings of this research study outline the potential for this approach. It is the practical application of the cognitive-model which makes research of this kind compelling and worthwhile to the educator within any sphere of professional development. As a practical

investigation of teacher-assessment, this study has added to the understanding of the newly important and expanding field of research into professional judgement.

BIBLIOGRAPHY

CITED REFERENCES

AHMANN J.S. and GLOCK M.D. (1975)
"Measuring and evaluating educational achievement"
BOSTON (Mass): ALLYN & BACON

AIRASIAN P.W., MADAUS G.F. and WOODS E.M. (1975)
"Scaling attitude items: a comparison of scalogram analysis and ordering theory"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 35, p809-819

ALLANSON J., KAVANAGH D. AND THOMAS N. (1990)
"Assessment and the National Curriculum: the standing of teachers and children"
THE CURRICULUM JOURNAL Vol 1 (2), p129-137

ARCHER J. and McCARTHY B. (1988)
"Personal biases in student assessment"
EDUCATIONAL RESEARCH Vol 30 (2), p142-145

ATKINSON L. (1990)
"The development of assessment procedures to accommodate the national curriculum, GCSE coursework and records of achievement within the mathematics faculty at Bridlington School 1984/89"
UNPUBLISHED MA DISSERTATION UNIVERSITY OF HULL

BARNES J.A. (1987)
"Setting directions for the 1990's"
PAPER TO AMERICAN VOCATIONAL ASSOCIATION LAS VAGAS (NV)

BART W.M. and KRUS D.J. (1973)
"An ordering-theoretic method to determine hierarchies among items"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 33, p291-300

BELL A. (1977)
"The APU and the 1978 mathematics survey"
MATHEMATICS TEACHING Vol 80, p24-27

BENNETT N.D. (1991)
"Change and continuity in school practice: a study of the influence affecting secondary school teachers' work and of the role of local and national policies within them"
UNPUBLISHED PhD THESIS BRUNEL UNIVERSITY

BLACK P. (1984)
"Standards of performance - expectations and reality"
JOURNAL OF CURRICULUM STUDIES Vol 16 (1), p94-96

BORG W.R and GALL M.D. (1983)
"Educational research: an introduction"
NEW YORK: LONGMAN

BORKO H. and CADWELL J. (1982)
"Individual differences in teachers' decision strategies: an investigation of classroom organization and management decisions"
JOURNAL OF EDUCATIONAL PSYCHOLOGY Vol 74 (4), p598-610

BROADFOOT P. (1980)
"Time for change: the problem of assessment"
FORUM FOR THE DISCUSSION OF NEW TRENDS IN EDUCATION Vol 23 (1), p18-20

BROADFOOT P. (1991)
"The conduct and effectiveness of primary education"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION CHICAGO (IL)

BROWN M. (1989)
"Graded assessment and learning hierarchies in mathematics - an alternative view"
BRITISH EDUCATION RESEARCH JOURNAL Vol 15 (2) p121-128

BUCKLE C.F. and RIDING R.J. (1988)
"Current problems in assessment - some reflections"
EDUCATIONAL PSYCHOLOGY Vol 8 (4), p299-306

CADWELL J. and JENKINS J. (1986)
"Teachers' judgements about their students: the effect of cognitive-simplification strategies on the rating process"
AMERICAN EDUCATIONAL RESEARCH JOURNAL Vol 23 (3), p460-475

CAMPBELL D.T. and STANLEY S.J. (1966)
"Experimental and quasi-experimental designs for research"
CHICAGO: RAND McNALLY

CAPLAN A. and McAFEE O. (1977)
"Classroom developmental assessment: the link between testing and teaching, an interim report"
INTERNATIONAL TRAINING CONSULTANTS DENVER (CO)

CARVER R.P. (1974)
"Two dimensions of tests: psychometric and edumetric"
AMERICAN PHYSCOLOGIST Vol 29, p512-518

CHARD S.C. (1990)
"The national curriculum of England and Wales: its implementation and evaluation in early childhood"
ERIC REFERENCE 337300

CHASE C.I. (1978)
"Measurement for educational evaluation"
READING (MASS): ADDISON-WESLEY

COCKROFT W.H. (1982)
"Mathematics counts"
LONDON: HMSO

COOPER H, FINDLEY M. and GOOD T. (1982)
"Relations between student achievement and various indexes of teacher expectation"
JOURNAL OF EDUCATIONAL PSYCHOLOGY Vol 74 (4), p577-579

CROSKERY K.M. (1988)
"curriculum based assessment in high school general mathematics classes"
DISSERTATION ABSTRACTS 8721356 UNIVERSITY OF NEVADA (NV)

CUTTANCE P. (1991)
"Monitoring educational quality through performance indicators for school practice: paper presented at conference"
AMERICAN EDUCATIONAL RESEARCH ASSOCIATION CHICAGO (IL)

DEALE R.N. (1976)
"Assessment and evaluation - neglected areas of inservice evaluation"
BRITISH JOURNAL OF INSERVICE EDUCATION Vol 2 (3) p204-208

DEPARTMENT OF EDUCATION AND SCIENCE (1988)
"National curriculum: task group on assessment and testing"
LONDON: HER MAJESTY'S STATIONARY OFFICE

DEPARTMENT OF EDUCATION AND SCIENCE (1989)
"Mathematics for ages 5 to 16"
LONDON: HER MAJESTY'S STATIONARY OFFICE

DEPARTMENT OF EDUCATION AND SCIENCE (1990)
"A summary of messages from recent speeches to teacher associations by the Rt. Hon. John MacGregor OBE MP, secretary of state for education and science"
LONDON: HER MAJESTY'S STATIONARY OFFICE

DESFORGES C. (1989)
"Testing and assessment"
LONDON: CASSELL

DRIVER R. and WORSLEY C. (1979)
"The assessment of performance in science project"
EUROPEAN JOURNAL OF SCIENCE EDUCATION Vol 1 (4), p441-447

DUSEK J.B. and JOSEPH G. (1983)
"The biases of teacher expectencies: a meta-analysis"
JOURNAL OF EDUCATIONAL PSYCHOLOGY Vol 75 (1), p327-346

EDWARDS A.L. (1960)
"Experimental design in pyschological research"
NEW YORK: HOLT, RIENHART and WINSTON

EGGLESTON J.F. (1979)
"Evaluating teachers or teaching?"
FORUM FOR THE DISCUSSION OF NEW TRENDS IN EDUCATION Vol 21 (2), p40-42

ERSKINE S.C. (1987)
"Centralism and professionalism: an inverse relationship?"
UNPUBLISHED MEd THESIS DUNDEE UNIVERSITY

FERRARA S.F. and THORNTON S.J. (1986)
"Using NEAP for state-by-state comparisons: the beginnings of a national achievement test and national curriculum: in the Nation's Report Card"
ERIC REFERENCE 279675

FINDLAY J. (1987)
"Criteria-based assessment in Queensland"
AUSTRALIAN MATHEMATICS TEACHER Vol 43 (3) p4-6

FITZGIBBON T.J. (1972)
"Norm referenced and criterion referenced tests form a publisher's point of view"
PAPER TO AMERICAN PSYCHOLOGICAL ASSOCIATION HONOLULU (HW)

FOXMAN D. and MITCHELL P. (1983)
"Assessing mathematics: 1. APU framework and modes of assessment"
MATHEMATICS IN SCHOOL Vol 12 (5), p2-5

FREMER J. (1973)
"Criterion-referenced interpretations of survey achievement tests"
EDUCATIONAL TESTING SERVICES PRINCETON (NY)

GAGNE R.M. (1968)
"Learning hierarchies"
EDUCATIONAL PSYCHOLOGIST Vol 6 (1) p3-6

GAIM TEAM (1988)
"The GAIM development pack"
DEVON: MACMILLAN EDUCATION

GALTON M. (1979)
"A constructive response to the APU"
FORUM FOR THE DISCUSSION OF NEW TRENDS IN EDUCATION Vol 22 (1) p20-22

GIPPS C. and GOLDSTEIN H. (1984)
"Local and national testing in the UK: the last ten years"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION NEW ORLEANS (LA)

GIPPS C. and GOLDSTEIN H. (1989)
"A curriculum for teachers assessment"
JOURNAL OF CURRICULUM STUDIES Vol 21 (6) p561-565

GIPPS C. (1990)
"Assessment: a teachers' guide to the issues"
LONDON: HODDER & STOUGHTON

GIPPS C. (1992)
"National testing at seven: what can it tell us?"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION SAN FRANSISCO (CA)

GLASER R. (1963)
"Instructional technology and the measurement of learning outcomes: some questions"
AMERICAN PSYCHOLOGIST Vol 18, p519-521

GLASER R. (1968)
"Evaluation of instruction and changing educational models"
CALIFORNIA UNIVERSITY CENTER FOR THE STUDY OF EVALUATION LOS ANGELES (CA)

GLASER R. and NITKO A.J. (1971)
"Measurement in learning and instruction. In Thorndike R.L., Educational measurement"
WASHINGTON: AMERICAN COUNCIL ON EDUCATION

GOLDSTEIN H. and NUTTALL D. (1985)
"Recent developments in assessment procedures in England and Wales"
PAPER TO NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION CHICAGO (IL)

GOODING C.T. (1980)
"An American looks at teacher views of the APU"
FORUM FOR THE DISCUSSION OF NEW TRENDS IN EDUCATION Vol 23 (1) p9-11

GRAY W.M. (1978)
"A comparison of piagetian theory and criterion-referenced measurement"
REVIEW OF EDUCATIONAL RESEARCH Vol 48 (2), p223-249

GREENEN J.P. and SMITH B.B. (1981)
"Assessing the generalizable skills of post-secondary vocational students: a validation study"
MINNEAPOLIS DEPARTMENT OF VOC. AND TECH. EDUCATION MINNESOTA UNIVERSITY

GREENEN J.P. (1984)
"The development of strategies and procedures for assessing the generalizable skills of students in secondary vocational programs"
URBANA DEPARTMENT OF VOC. AND TECH. EDUCATION ILLINOIS UNIVERSITY

GULLO D.F. and AMBROSE R.P. (1987)
"Perceived competence and social acceptance in kindergarten: its relationship to academic performance"
JOURNAL OF EDUCATIONAL RESEARCH Vol 8 (1), P28-32

HART K. (1980)
"Secondary school children's understanding of mathematics, report of the mathematics component of the Concepts in secondary mathematics and science programme"
UNIVERSITY OF LONDON: CHELSEA COLLEGE

HARTNETT A. and NAISH M. (1990)
"The sleep of reason breeds a monsters: the birth of a statutory curriculum in England and Wales"
JOURNAL OF CURRICULUM STUDIES Vol 22 (1), p1-16

HAYES S. (1991)
"Too eagerly awaited assessment?"
BRITISH JOURNAL OF SPECIAL EDUCATION Vol 18 (2), p48-51

HENRY T.B. (1991)
"Spotlight of a century of educational reform"
ERIC REFERENCE 343247

HER MAJESTY'S INSPECTORATE (1989)
"The implementation of the national curriculum in primary schools: a survey of 500 schools"
LONDON: HER MAJESTY'S STATIONARY OFFICE

HER MAJESTY'S INSPECTORATE (1990)
"The implementation of the national curriculum in primary schools: a survey of 100 schools"
LONDON: HER MAJESTY'S STATIONARY OFFICE

HER MAJESTY'S INSPECTORATE (1991)
"Mathematics key stages 1 and 3: a report on the first year, 1989-90"
LONDON: HER MAJESTY'S STATIONARY OFFICE

HOGE R.D. and BUTCHER R. (1984)
"Analysis of teacher judgements of pupil achievement levels"
JOURNAL OF EDUCATIONAL RESEARCH Vol 76 (5), p777-781

HOSTE R. and BLOOMFIELD B. (1975)
"Continuous assessment in the CSE: opinion and practice"
LONDON: EVANS/METHUEN

ISLEY M.M. (1989)
"Teachers as effective curriculum implementers: evaluation of an experiment in in-service education"
UNPUBLISHED MEd THESIS UNIVERSITY OF LEICESTER

INGVASSON L. (1990)
"Enhancing professional skill and accountability in the assessment of student learning"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION BOSTON (MA)

JARMAN R. (1990)
"Primary science - secondary science community: a new era?"
SCHOOL SCIENCE REVIEW Vol 71 (257), p19-29

JASMAN A.M. (1987)
"Teacher based assessments: a study of development, validity and reliability of teachers' assessments"
UNPUBLISHED PhD THESIS UNIVERSITY OF LEICESTER

JOHNSON S. (May 1989)
"Beloe to Baker: thirty years of teacher assessment and moderation"
MIDLAND EXAMINING GROUP Research paper No. 1

KINGDON J.M. and HARTLEY D.J. (1982)
"Teacher assessment of University of London A-level biology practical notebooks - a report on the first operational examination"
JOURNAL OF BIOLOGICAL EDUCATION Vol 16 (4), p286-292

LAWRENCE G. (1974)
"Inservice training - what the teachers want"
BRITISH JOURNAL OF INSERVICE EDUCATION Vol 1 (2) p44-53

LINDVALL C.M. and NITKO A.J. (1969)
"Criterion-referenced testing and the individualization of instruction"
PAPER TO NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION LOS ANGELES (CA)

LINDVALL C.M. and NITKO A.J (1975)
"Measuring pupil achievement and aptitude"
NEW YORK: HARCOURT BRACE

LOFTY J.S. (1990)
"Coming home to the national curriculum (news from England)"
ENGLISH EDUCATION Vol 22 (4), p241-264

LONGSTAFF S.A. (1990)
"Democracy and the idea of a compulsory national curriculum: a philosophical examination"
ASLIB ABSTRACTS 39-5928 CAMBRIDGE UNIVERSITY

LORD J. (1987)
"School based assessment part 3:some practical problems"
MATHEMATICS IN SCHOOL Vol 16 (5), p9-10

LUIJTEN A.J.M. (1991)
"Issues in public examinations: a selection of proceedings"
IAEA CONFERENCE MAASTRICHT NETHERLANDS

MARAJORAM D.T.E. (1978)
"The APU and assessment in the midlle years"
EDUCATION 3-13 Vol 6 (2), p31-36

McBRYDE B. and LAMONT A. (1980)
"Mapping of assessment procedures in schools: a study of assessment procedures in three subject areas at year 12 level in Queensland State secondary schools"
ERIC REFERENCE 205569

McGUINESS P.J. (1985)
"Syllabus development and examination reform: a case study in the development in post primary mathematics in Northern Ireland"
UNPUBLISHED PhD THESIS UNIVERSITY OF DUBLIN

MICHAEL M. (1987)
"Individual learning and differentiation: more difficult than it seems"
ASLIB ABSTRACTS 38-2678 ABERDEEN UNIVERSITY

MOBLEY M., EMERSON C., GODDARD I., GOODWIN S. and LETCH R. (1986)
"All about GCSE: a clear and concise summary of all the basic information about GCSE"
LONDON: HEINEMANN

MOORE C.H. (1989)
"An analysis of teacher and pupil perceptions to the introduction of the school-based-assessment of practical work in GCSE chemistry in Northern Ireland"
ASLIB ABSTRACTS 39-5756 QUEEN'S UNIVERSITY BELFAST

MURPHY P. (1988)
"Insights into pupils' responses to practical investigations from the APU"
PHYSICS EDUCATION Vol 23 (6), p330-336

NELSON G.H. (1988)
"The introduction of mathematics coursework into Lincolnshire secondary schools"
ASLIB ABSTRACTS 39-8272 UNIVERSITY OF LOUGHBOROUGH

NITKO A.J. (1980)
"Distinguishing the many varieties of criterion-referenced tests"
REVIEW OF EDUCATIONAL RESEARCH Vol 50 (3) p461-485

NOSS R., GOLDSTEIN H. and HOYLES C. (1989)
"Graded assessment and learning hierarchies in mathematics"
BRITISH EDUCATION RESEARCH JOURNAL Vol 15 (2) p109-120

NUTTALL D.L. and WILLMOTT A.S. (1972)
"British exams: techniques of analysis"
SLOUGH: NATIONAL FOUNDATION FOR EDUCATIONAL RESEARCH

NUTTALL D. (1992)
"Linkages between new criteria and curriculum development"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION SAN FRANCISCO (CA)

OSBORN M. (1991)
"The impact of current changes in English primary schools on teacher professionalism"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION CHICAGO (IL)

OWEN S.A. (1976)
"The validity of student ratings: a critique"
STORES BUREAU OF EDUCATIONAL RESEARCH AND SERVICE CONNECTICUT UNIVERSITY (MA)

PAGE B. and HEWETT D. (1987)
"Languages step by step: graded objectives in the UK"
CENTRE FOR INFORMATION ON LANGUAGE TEACHING AND RESEARCH LONDON

PEDULLA J.J., AIRASIAN P.W. and MADAUS G.F. (1980)
"Do teacher ratings and standardised test results of students yield the same information?
AMERICAN EDUCATIONAL RESEARCH Vol 17 (3) p303-307

PENNYCUICK D.B. (1987)
"The development, use and impact of graded tests, with particular reference to modern languages, mathematics and science"
UNPUBLISHED PhD THESIS UNIVERSITY OF SOUTHAMPTON

PENNYCUICK D.B. and MURPHY R.J. (1986)
"Mastery validity and comparability issues in relation to graded assessment schemes"
STUDIES IN EDUCATIONAL EVALUATION Vol 12 (3), p305-311

PIKE M. and MURRAY L. (1991)
"Assessing open ended tasks"
MATHEMATICS IN SCHOOL Vol 20 (2), p32-33

PIRIE S. (1988)
"GCSE coursework: mathematics"
BASINGSTOKE: MACMILLAN EDUCATION

POPHAM W.J. and HUSEK T.R. (1969)
"Implications of criterion-referenced measurement"
JOURNAL OF EDUCATIONAL MEASUREMENT Vol 6, p1-9

POPHAM W.J. (1978)
"Criterion-referenced measurement"
NEW JERSEY: PRENTICE-HALL

POPHAM W.J. (1981)
"Modern educational meausrements"
NEW JERSEY: PRENTICE-HALL

PORTAL M. (1991)
"Graded assessment: an assessment model for key stage reporting"
BRITISH JOURNAL OF CURRICULUM AND ASSESSMENT Vol 1 (3), p21-24

PRESTON M. (1980)
"The first APU primary maths survey: an appraisal"
EDUCATION 3-13 Vol 8 (2), p35-40

RABAN B. (1991)
"English assessment in a national curriculum"
INTERNATIONAL READING ASSOCIATION LAS VAGAS (NV)

RADNOR H.A. (1991)
"A moderation model for pupils' teacher-assessed work"
BRITISH JOURNAL OF CURRICULUM AND ASSESSMENT Vol 1 (3), p15-20

RELF S. (1990)
"The story of Frank, aged eleven"
MATHEMATICS IN SCHOOL Vol 19 (2), p40-41

ROBINSON C.G. (1988)
"Assessment and the curriculum"
EDUCATIONAL PSYCHOLOGY: AN INTERNATIONAL JOURNAL OF EXPERIMENTAL EDUCATIONAL PSYCHOLOGY Vol 8 (4), p221-227

SCHOOL EXAMINATION AND ASSESSMENT COUNCIL (Spring 1991)
"SEAC recorder: bulletin No. 7"
LONDON: HER MAJESTY'S STATIONARY OFFICE

SCHOOLS COUNCIL (1967a)
"Standards in CSE and GCE: English and mathematics: working paper No. 9"
LONDON: HER MAJESTY'S STATIONARY OFFICE

SCHOOLS COUNCIL (1967b)
"The Certificate of Secondary Education trial examinations: written English examinations bulletin No. 16"
LONDON: HER MAJESTY'S STATIONARY OFFICE

SCHOOL MATHEMATICS PROJECT (1990)
"Assessing mathematics within the national curriculum: a resource bank for teachers"
SMP OFFICE SOUTHAMPTON UNIVERSITY

SCHRODER C. and CRAWFORD P. (1970)
"School achievement as measured by teacher ratings and standardized achievement tests"
TORONTO BOARD OF EDUCATION RESEARCH DEPARTMENT TORONTO (ON)

SECONDARY EXAMINATIONS COUNCIL (1986)
"School based assessment"
WORKING PAPER 3 SECONDARY EXAMINATIONS COUNCIL

SHARP S. (1991)
"Curriculum and assessment in scotland 5-16 part I: 14-16 developments"
BRITISH JOURNAL OF CURRICULUM AND ASSESSMENT Vol 1 (2), p15-16

SHALVERSON R.J. and STERN P. (1981)
"Research on teachers' pedagogical thoughts, judgements, decisions and behaviour"
REVIEW OF EDUCATIONAL RESEARCH Vol 51 (4), p455-498

SHAYER M. (1991)
"Improving standards and the national curriculum"
SCHOOL SCIENCE REVIEW Vol 72 (260), p17-24

SLOVIC P. and LICHTENSTEIN S. (1971)
"Comparison of bayesian and regression approaches to the study of information processing in judgement"
ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE Vol 6, p649-744

STONES E. (1979)
"The world of APU"
FORUM FOR THE DISCUSSION OF NEW TRENDS IN EDUCATION Vol 22 (1), p12-13

SWAIN J.R. (1991)
"The nature and assessment of scientific explorations in the classroom"
SCHOOL SCIENCE REVIEW Vol 72 (260), p65-77

TAMIR P. (1987)
"Testing and the school curriculum: evolving trends"
STUDIES IN EDUCATIONAL EVALUATION Vol13 (1), p3-6

THE LOW ATTAINERS IN MATHEMATICS PROJECT TEAM. (1987)
"Better mathematics: a curriculum development study based on the low attainers in mathematics project"
LONDON: HER MAJESTY'S STATIONARY OFFICE

THORNDIKE R.L. and HAGEN E.P. (1977)
"Measurement and evaluation in psychology and education"
NEW YORK: WILEY

TORRÂNCE H. (1986)
"Expanding school based assessment: issues, problems and future possibilities"
RESEARCH PAPERS IN EDUCATION Vol 1 (1), p48-59

TORRÂNCE H. and MURPHY R. (1988)
"The changing face of educational assessment"
MILTON KEYNES: OPEN UNIVERSITY PRESS

TROMAN G. (1989)
"Testing tensions: the politics of educational assessment"
BRITISH EDUCATION RESEARCH JOURNAL Vol 15 (3), p279-295

VAN DER KAMP L.J.T. (1976)
"Agreement between raters"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 36, p311-317

VERGNAUD G. (1990)
"Mathematics and cognition: a research synthesis by the international group for the psychology of mathematics education"
CAMBRIDGE: CAMBRIDGE UNIVERSITY PRESS

WAKEFIELD J.A. (1980)
"The relationship between two expressions of reliability: percentage agreement and phi"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 40, p593-597

WHEELER A.H. (1985)
"Beyond the crossroads: charting the future"
PRESENTED TO COLLEGE OF EDUCATION AND BEHAVIORAL SCIENCE, MISSOURI STATE UNIVERSITY (MS)

WILLIAM D. (1992a)
"Some technical issues in assessment: a user's guide"
BRITISH JOURNAL OF CURRICULUM AND ASSESSMENT Vol 2 (3), p11-20

WILLIAM D. (1992b)
"INSET for national curriculum assessment: lessons from the key stage 3 SATs trials and pilot"
BRITISH JOURNAL OF CURRICULUM AND ASSESSMENT Vol 2 (2), p8-11

WRIGHT D. and WIESE J. (1988)
"Teacher judgements in student evaluation: a comparison of grading methods"
JOURNAL OF EDUCATIONAL RESEARCH Vol 82 (1), p10-14

OTHER REFERENCES

ABDULLAH K.B and LOVELL W.E. (1981)
"A scalogram analysis of two measures of concept generalizability"
PAPER TO NATIONAL ASSOCIATION FOR RESEARCH IN SCIENCE TEACHING ELLENVILLE (NY)

BADIAN N. A. (1976)
"Early prediction of academic underachievement"
PAPER TO THE COUNCIL FOR EXCEPTIONAL CHILDREN CHICAGO (IL)

BAIN D.M. (1988)
"School based assessment part 4: the assessment of group work"
MATHEMATICS IN SCHOOL Vol 17 (1), p10-11

BENEFIELD K. and CAPIE W. (1976)
"An empirical derivation of hierarchies of propositions related to ten of Piaget's sixteen binary operations"
JOURNAL OF RESEARCH IN SCIENCE TEACHING Vol 13 (3), p193-204

BOCK R.D. and MISLEVY R. (1986)
"Comprehensive educational assessment for the States"
CENTER FOR THE STUDY OF EVALUATION CALIFORNIA UNIVERSITY (CA)

BRAUND M. (1990)
"Be able to....what?"
SCHOOL SCIENCE REVIEW Vol72 (259), p57-62

BROADFOOT P. (1980)
"Time for change the problem of assessment"
FORUM FOR THE DISCUSSION OF NEW TRENDS IN EDUCATION Vol 23 (1), p18-20

CAPIE W. and JONES H.L. (1971)
"An assessment of hierarchy validation techniques"
JOURNAL OF RESEARCH IN SCIENCE TEACHING Vol 8 (2), p137-147

COHEN L. and HOLLIDAY M. (1979)
"Statistics for education and physical education"
LONDON: HARPER & ROW

CONNER J.E. and LESSINGER L.M. (1976)
"Educational auditing and quality assurance: occasional paper No. 4"
COUNCIL OF CHIEF STATE SCHOOL OFFICERS WASHINGTON DC

CZIKO G. A. (1984)
"An improvement over Guttman scalogram analysis: a computer program for evaluating cumulative, nonparametric scales of dichotomous items"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 44 (1), p159-163

DAYTON C.M. and MaCREADY G.B. (1976)
"A probabalistic model for validation of behavioral hierarchies"
PSYCHOMETRIKA VoL 41 (2), p180-204

FAIR P.C. (1986)
"A consideration of some aspects of professional and govermental control of the curriculum"
ASLIB ABSTRACTS 37-369 UNIVERSITY OF LEICESTER

FUENTES E.J. and WISENBAKER J.M. (1979)
"The use of teacher rating of oral English proficiency as a covariate in the analysis of reading scores"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION SAN FRANCISCO (CA)

GIPPS C. (1989)
"A curriculum for teacher assessment"
JOURNAL OF CURRICULUM STUDIES Vol 21 (6), p561-565

GRAHAM H. (1989)
"Teachers' assessment of practical skills in biology"
ASLIB ABSTRACTS 38-7040 UNIVERSITY OF NOTTINGHAM

GRIFFITHS S. (1988)
"GCSE: report of the working party on English Literature"
USE OF ENGLISH Vol 40 (1), p19-36

GUAY R.B. and McCABE G.P. (1978)
"A chi-square test for hierachical dependency"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION TORONTO (ON)

HADLEY M. and VITALE P. (1985)
"Evaluating student achievement"
ERIC REFERENCE 285878

HANNAN B. (1985)
"Assessment and evaluation in schooling"
VICTORIA: DEAKIN UNIVERSITY PRESS (AUS)

HSU T.C. (1971)
"Empirical data on criterion referenced tests"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION NEW YORK (NY)

JARVIS P. (1986)
"Plowden, Piaget and the secondary science curriculum"
ASLIB ABSTRACTS 36-7268 UNIVERSITY OF EXETER

JOHNSON S.C. (1967)
"Hierarchical clustering schemes"
PSYCHOMETRIKA Vol 32 (3), p241-254

KAMBOURI M. (1991)
"Knowledge assessment: a comparison between human experts and computerised procedures"
DISSERTATION ABSTRACTS 9213243 NEW YORK UNIVERSITY (NY)

LIDSTONE P. (1991)
"Strategies for the teacher-assessed component of key stage 3 science"
SCHOOL SCIENCE REVIEW Vol 72 (260), p137-141

MASTERS G.N. and EVANS J. (1986)
"A sense of direction in criterion-referenced assessment"
STUDIES IN EDUCATIONAL EVALUATION Vol12 (3), p257-265

MADAUS G.F. (1985)
"Public policy and the testing profession"
EDUCATIONAL MEASUREMENT: ISSUES AND PRACTICE Vol 4 (4), p5-11

MANOS K. (1987)
"The threat to mass testing: how do we prepare?"
PAPER TO CONFERENCE ON COLLEGE COMPOSITION AND COMMUNICATION ATLANTA (GA)

MASTERS G.N. (1984)
"DICOT: analyzing classroom tests with the Rasch model"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 44 (1), p145-150

McNAMARA D. (1990)
"The national curriculum: an agenda for research"
BRITISH EDUCATION RESEARCH JOURNAL Vol 16 (3), p225-235

McQUITTY L.L. and KOCH V.L. (1976)
"Highest column entry hierarchical clustering: a redevelopment and elaboration of elementary linkage analysis"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 36 (2), p243-258

MILTON K.G. (1985)
"Problem solving in mathematics teaching"
JOURNAL OF SCIENCE AND MATHEMATICS EDUCATION IN SOUTH EAST ASIA Vol 8 (1), p7-10

NUTTALL D. (1988)
"The implications of national curriculum assessments"
EDUCATIONAL PYSCHOLOGY: AN INTERNATIONAL JOURNAL OF EXPERIMENTAL EDUCATIONAL PSYCHOLOGY Vol 8 (4), p229-236

POSTLETHWAITE T.N. (1986)
"The use of standardized tests in secondary schools in four european countries"
NATIONAL CENTER ON EFFECTIVE SECONDARY SCHOOLS MADISON (WS)

REED L. (1983)
"Assessing children's speaking, listening, and writing skills"
DINGLE ASSOCIATES INC. WASHINGTON DC

REID J.B. and ROBERTS D.M. (1978)
"A Monte Carlo comparison of phi and kappa as measures of criterion-referenced reliability"
PAPER TO AMERICAN EDUCATIONAL RESEARCH ASSOCIATION TORONTO (ON)

ROSS J.A. and MAYNES F.J. (1983)
"Development of a test of experimental problem-solving skills"
JOURNAL OF RESEARCH IN SCIENCE TEACHING Vol 20 (1) p63-75

SCHOOLS COUNCIL (1967c)
"Teachers' experience of school based examining (English and physics): written English examinations bulletin No. 15"
LONDON: HER MAJESTY'S STATIONARY OFFICE

SCHOOLS COUNCIL (1974)
"Dissemination and in-service training: report of the Schools Council Dissemination Working Party: pamphlet No. 14"
LONDON: HER MAJESTY'S STATIONARY OFFICE

SCOTTISH COUNCIL FOR RESEARCH IN EDUCATION (1982)
"The Scottish Council for research in education: fifty-fourth annual report 1981-82"
SCOTTISH COUNCIL FOR RESEARCH IN EDUCATION EDINBURGH

SMITH J.K. (1980)
"On the examination of test unidimensionality"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 40 (4), p885-889

TAYLOR R.M. (1990)
"The national curriculum: a study to compare levels of attainment with data from APU science surveys (1980-84)"
SCHOOL SCIENCE REVIEW Vol 72 (258), p31-37

THOMAS D. (1989)
"The national curriculum - moving into focus?"
PAPER TO LEAU NATIONAL CONFERENCE LIVERPOOL

TORRANCE H. (1986)
"Assessment and examinations: social context and educational practice"
ASLIB ABSTRACTS 36-7282 UNIVERSITY OF EAST ANGLIA

WATKINSON A. (1991)
"Primarily assessing"
EDUCATION IN SCIENCE Vol 10 (141), p8-9

WEBBER M.B. (1977)
"An examination of the bilevel dimensionality of lower and higher mental processes in probability achievement"
EDUCATIONAL AND PYSCHOLOGICAL MEASUREMENT Vol 37 (4), p987-990

WILSON C. (1987)
"Better science: assessing progress. Curriculum guide 11"
SCHOOL CURRICULUM DEVELOPMENT COMMITTEE LONDON

WIMBERLEY R.C. (1976)
"ALAM and ALAS: questioning error assignments in unidimensional Guttman scaling"
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT Vol 36 (2), p361-367

WOODROW D. (1992)
"Learning from experience: some principles of INSET practice"
MATHEMATICS IN SCHOOLS Vol 20 (4), p11-13

YOUNGMAN M.B. and EGGLESTON J.F. (1979)
"Constructing tests and scales: rediguide 10"
UNIVERSITY OF NOTTINGHAM SCHOOL OF EDUCATION NOTTINGHAM

LIST OF APPENDICES

1. Pre-pilot questionnaire, including an example of actual responses.

2. Information regarding the Grantley and York training courses (North Yorkshire).

3. Postal questionnaire address list (stratified sample of Humberside secondary schools).

4. Letter to heads of mathematics within the main study (postal questionnaire).

5. Mathematics Attainment Targets within the National Curriculum.

6. Example of resources consulted during the hypothetical profile construction aspect of the research.

7. Test Specifications: Introduction.

8. Test Specifications: Document A (including annotation by a validation judge).

9. Test Specifications: Document B.(including annotation by a validation judge).

10. Test Specifications: Document C (an example of a validation judge's summary decisions and recommendations).

11. Grantley/York 'flysheet' questionnaire, including an example of actual responses.

12. 'Pull-out' reduced questionnaire: phase I (version 1).

13. 'Pull-out' reduced re-test questionnaire: phase I.

14. 'Pull-out' reduced 'flysheet' of hypothetical profiles with informational cue status (not included in the actual test-instrument version).

15. 'Pull-out' reduced questionnaire: phase III (version 3).

16. Full size example of phase III (version 3) questionnaire (within binding back pocket).

17. Activity sampling results (Humberside County Council) consulted during the procedural aspect of research.

18. INSET package (York venue - phase II) used in conjunction with version 2 of the test-instrument.

Questionnaire re: National Curriculum Assessment.

The questions outlined below are part of a pilot study into the actual and potential problems encountered by teachers performing Teacher Assessments, both now and in the future, and as such your responses will form the basis of further work: As a consequence, many of the questions are open ended. I would be grateful if you could fill in each section using the spaces provided. If you need more space feel free to continue on a separate sheet.

Section 1. - personal details (tick where appropriate)

1. Number of years teaching mathematics 12

2. Is this your principle subject YES ✓ NO __ ? If NO, what is?

3. Previous experience of criterion referenced assessment YES __ NO ✓?
   If YES please specify details.

4. Have you been involved in Teacher Assessment since Sept'89 YES ✓ NO __?
   If YES please specify details.
   SMP 11-16 Booklets and related work
   Assessment as part of record of achievement assessing topics, skills, knowledge immediately after work on a topic. Not including assessment of sustained skills with regard to Nat. Curr.

Section 2. - Statements of Attainment (SoA)

Please comment on the positive and negative aspects of using SoA for assessing a pupil's work. Make reference, if possible, to the quality, degree of difficulty and uniformity of SoA; and also your interpretation of them - is it narrow, broad or absolute.

1) Imprecise – they are capable of being interpreted in a range of ways - e.g. linear equations, area.
2) It is fairly easy for pupils to be assessed as 'knowing' items out level 5 say, without understanding at level 4 in the same AT e.g. Probability – pupils can learn the rules of tree diagrams (level 6) without understanding the idea of probability (level 3)
3) There is no 'match' of levels across different AT's

Section 3. - Pre-conceptions

In your opinion, do you feel you have pre-conceived ideas of a pupil's level of attainment due to neatness of work, language used, a their group or set etc. If yes, please specify; how do you compensate for these pre-conceptions during any assessments you undertake?

Yes, it is impossible to avoid preconceptions.
In truth we tend not to compensate since we assess by differentiating by task.
However, the preconceptions are reduced with the length of time you have taught each pupil. They are also evened out by changes of teacher. Undoubtedly some pupils still suffer or gain

Section 4. - Sustainability

Please give your interpretation of the term sustainability in the context of assessment. If possible relate this to the terms Knowledge, Skill and Understanding.

Using our system, once an SoA is awarded, it cannot be taken away. I have serious doubts that many pupils would be awarded the same SoA in 6 months time. Their skills and understanding are being built on all the time, but actual knowledge can be lost. Re-inforcement is a major part of the course so there should not be any real problem.

Section 5. - Levels and Examples

(a) Comment on the comparability of levels across the 14 ATs, make reference to their degree of difficulty and uniformity of composition.

I believe there is some lack of uniformity across the ATs. Analysis would be a lengthy process.

(b) How important/helpful, or not, are the examples given in the National Curriculum folder (statutory orders). Make reference, if possible, to their influence and quality.

Examples do give a guideline as to the level of difficulty, but where more than one example is given e.g. 6/8a, there is not only a range of techniques necessary, but also discrepancies in the degree of difficulty.

Section 6. - Important Issues

Please outline what you consider to be the most important issue/concern with regard to Teacher Assessment of pupil's work using SoAs and National Curriculum Levels.

Time ; lack of uniformity of approach by individuals ; parental confusion ; employer's confusion ;

Thank you for your cooperation and time,

Les Atkinson - Dec '90.

Training for National Curriculum Assessment - Key Stage 3

Monday 4th November 1991
Grantley Hall
9.15 a.m. - 4.00 p.m.

Part 3

PROGRAMME

9.15 a.m.	Arrival
9.30 - 9.45	Developing Teachers' Professional Judgement - L. Atkinson, Head of Mathematics, Pindar School.
9.45 - 11.00	Mental Mathematics and Computation.
	Coffee
11.15 - 12.30 p.m.	Feedback from the Newspaper and Rolling Ball activities.
1.45 - 2.45	Developing children's awareness of their mathematical ability (2) - Developing Strategies and Reasoning.
2.45 p.m.	Tea
3.00 - 3.30	Classroom activities in preparation for Part 4 of the course.
3.30 - 4.00	Continuation of Teachers' Professional Judgement.
	Departure.

Please bring with you,

1. The National Curriculum - Mathematics 5 to 16/1991.(Proposals)
2. The report of and be prepared to discuss the classroom activities from Part 2 of the course.
3. The three pieces of children's work for the Newspaper and the Rolling Ball activities.

Please complete and return the confirmation slip below by Friday 18th October, 1991, to P. J. Wells Inspector/Advisers' Office, White Cross Lodge, 150 Haxby Road, York YO3 7JN.

- -

To: Mr P. J. Wells, Senior Inspector/Adviser (Mathematics) White Cross Lodge, 150 Haxby Road, York YO3 7JN.

I confirm that I will attend Part 3 of the Course at Grantley Hall on Monday 4th November, 1991.

Name _____

School _____

Special Diet _____

Thursday . 7th November. 1991.
York Staff Development Centre. Park Grove.
9.15 a.m. 4.00 p.m.

Part 3

PROGRAMME

9.15 a.m	Arrival
9.30 - 9.45	Developing Teachers' Professional Judgement - L. Atkinson, Head of Mathematics, Pindar School.
9.45 - 11.00	Mental Mathematics and Computation.
	Coffee
11.15 - 12.30 p.m.	Feedback from the Newspaper and Rolling Ball Activities.
1.45 - 2.45	Developing children's awareness of their mathematical ability (2) - Developing Strategies and Reasoning.
2.45 p.m.	Tea
3.00 - 3.30	Classroom activities in preparation for Part 4 of the course.
3.30 - 4.00	Continuation of Teachers' Professional Judgement.
	Departure.

Please bring with you,

1. The National Curriculum - Mathematics 5 to 16/1991.(Proposals)

2. The report of and be prepared to discuss the classroom activities from Part 2 of the course.

3. The three pieces of children's work for the Newspaper and the Rolling Ball Activities.

Please complete and return the confirmation slip below by Friday 18th October, 1991, to P. J. Wells Inspector/Advisers' Office, White Cross Lodge, 150 Haxby Road, York YO3 7JN.
PLEASE NOTE: Venue:
Staff Development Centre - formerly York Education, Park Grove, York.

- -

To: Mr P. J. Wells, Senior Inspector/Adviser (Mathematics) White Cross Lodge, 150 Haxby Road, York YO3 7JN.

I confirm that I will attend Part 3 of the Course at York Staff Development Centre, Park Grove, York on Thursday .7th November, 1991.

Name _____

School _____

Special Diet _____

PJWB013

(1) The Head of Mathematics,
Headlands School,
Sewerby Road,
Bridlington,
North Humberside.
YO16 5UR

(2) The Head of Mathematics,
Market Weighton School,
Spring Road,
Market Weighton,
North Humberside.
YO4 3JE

(3) The Head of Mathematics,
St. Mary's RC School,
Wooton Road,
Grimsby,
South Humberside.
DN33 1HE

(4) The Head of Mathematics,
Western School,
Cambridge Road,
Grimsby,
South Humberside.
DN34 5TE

(5) The Head of Mathematics,
David Lister School,
Rustenberg Street,
Hull,
North Humberside.
HU9 2PR

(6) The Head of Mathematics,
Sir Henry Cooper School,
Thorpepark Road,
Orchard Park Estate,
Hull,
North Humberside.
HU6 9ES

(7) The Head of Mathematics,
South Leys School,
Enderby Road,
Scunthorpe,
South Humberside.
DN17 2JL

(8) The Head of Mathematics,
Winterton Comprehensive,
Newport Drive,
Winterton,
Scunthorpe,
South Humberside.
DN15 9QD

(9) The Head of Mathematics,
Pocklington Woldgate School,
Kilnwick Road,
Pocklington,
North Humberside.
YO4 2LL

(10) The Head of Mathematics,
Withernsea High School,
Hull Road,
Withernsea,
North Humberside.
HU19 2EQ

(11) The Head of Mathematics,
Whitgift School,
Crosland Road,
Grimsby,
South Humberside.
DN37 9EH

(12) The Head of Mathematics,
Amy Johnson School,
Ringrose Street,
Hull,
North Humberside.
HU3 5QB

(13) The Head of Mathematics,
William Gee School,
Bishop Alcock Road,
Hull,
North Humberside.
HU5 4RS

(14) The Head of Mathematics,
Brumby Comprehensive,
Cemetery Road,
Scunthorpe,
South Humberside.
DN16 1NT

(15) The Head of Mathematics,
North Axholme Comprehensive,
Wharf Road,
Crowle,
Scunthorpe,
South Humberside.
DN17 4HU

(16) The Head of Mathematics,
Vale of Ancholme School,
Westmoor House,
Grammer School Road,
Brigg,
South Humberside.
DN20 8BA

(17) The Head of Mathematics,
Cottingham High School,
Harland Way,
Cottingham,
North Humberside.
HU16 5PX

(18) The Head of Mathematics,
Waltham Toll Bar School,
Station Road,
New Waltham,
Grimsby,
South Humberside.
DN36 4RZ

(19) The Head of Mathematics,
The Healing School,
Healing,
Nr. Grimsby,
South Humberside.
DN37 7QD

(20) The Head of Mathematics,
Kelvin Hall School,
Bricknell Avenue,
Hull,
North Humberside,
HU5 4QH

(21) The Head of Mathematics,
Sydney Smith School,
First Lane,
Anlaby,
Hull,
North Humberside.
HU10 6UU

(22) The Head of Mathematics,
High Ridge Comprehensive,
Doncaster Road,
Scunthorpe,
South Humberside,
DN15 7DF

(23) The Head of Mathematics,
St. Bede's RC School,
Collum Avenue,
Ashby,
Scunthorpe,
South Humberside.
DN16 2TF

(24) The Head of Mathematics,
Vermuyden School,
Centenary Road,
Goole,
Humberside.
DN14 6AN

(25) The Head of Mathematics,
South Holderness School,
Station Road,
Preston,
Hull,
North Humberside.
HU12 8UZ

(26) The Head of Mathematics,
Wintringham School,
Weelsby Avenue,
Grimsby,
South Humberside.
DN32 0AZ

(27) The Head of Mathematics,
Malet Lambert School,
James Reckitt Avenue,
Hull,
North Humberside.
HU8 0JD

(28) The Head of Mathematics,
Newland School,
Cottingham Road,
Hull,
North Humberside.
Hu6 7RU

(29) The Head of Mathematics,
Perronet Thompson School,
Wawne Road,
Bransholme,
Hull,
North Humberside.
HU7 4WR

(30) The Head of Mathematics,
Baysgarth School,
Barrow Road,
Barton-on-Humber,
South Humberside.
DN18 6AE

(31) The Head of Mathematics,
Snaith School,
Pontefract Road,
Snaith,
South Humberside.
DN14 9LB

(32) The Head of Mathematics,
South Axholme Comprehensive,
Burnham Road,
Epworth,
Doncaster,
South Humberside.
DN9 1BY

To the Head of Mathematics.

I would be most grateful if you could help me with a research project I am undertaking (at Hull University) into Teacher Assessments within the National Curriculum. The work is based on the old NC Statements of Attainment (SoAs) but the nature of the work makes it equally applicable to the new SoAs.

Your cooperation would require the involvement of up to 7 members of your department for 20 minutes at the most. The work is in the form of a three part questionnaire. The first part is concerned with personal details, the second and third parts involve the matching of pupils' work to Statements of Attainment - the very essence of Teacher Assessment.

I have enclosed the following items:

Seven A3 sheets (closed with a paperclip).

Seven A4 'flysheets' of pupils work.

S.A.E to return the completed A3 sheets to me by 20th Dec' 1991. (if possible)

N.B. you may keep the 'flysheets' (pages 2 & 4), I only need you to return the completed A3 sheets.

If you do decide to cooperate then I would appreciate it if you could coordinate the distribution and collection of these A3 sheets and 'flysheets' to your staff. Each member of staff requires an A3 sheet and an A4 'flysheet'. Each A3 sheet has a front cover with detailed instruction which they will need to follow very carefully.

In conclusion:

I have already used a similar version of this questionnaire and made a preliminary analysis of the results. Should you require the results of that and indeed the analysis of the responses to this questionnaire, then I will provide you with the details (I will need an S.A.E though, sorry). I anticipate the results will be ready by early February 1992. In any case, I will provide Peter Lacey with the results and he may or may not dispatch these to schools as a matter of course, or you could contact him direct for the information. The preliminary findings I have at the moment have been substantiated by at least one other group working independently on this problem. Unfortunately, to reveal even the preliminary findings at this stage would invalidate the purpose of this questionnaire - sorry! Finally, I believe the findings, when formulated properly, will be both informative and useful to you in the planning of your Teacher Assessments in the near future.

Thank you for your attention,

Les Atkinson - Head of Mathematics Pindar School.

N.B. - could you remind your staff that on completion of their questionnaire it is vital they do not return to previous parts and make any alterations - this will ensure the exercise is authentic.

Attainment Targets — Maths —

	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5	LEVEL 6
Attainment Target 1	• use materials provided for a task. • talk about own work and ask questions. • make predictions based on experience.	• select the materials and the mathematics to use for a task. • describe current work, record findings and check results. • ask and respond to the question: 'What would happen if . . .?'	• select the materials and the mathematics to use for a task; check results and consider whether they are sensible. • explain work being done and record findings systematically. • make and test predictions.	• select the materials and the mathematics to use for a task; plan work methodically. • record findings and present them in oral, written or visual form as appropriate. • use examples to test statements or definitions.	• select the materials and the mathematics to use for a task; check there is sufficient information; work methodically and review progress. • interpret mathematical information presented in oral, written or visual form. • make and test simple statements.	• design a task and select appropriate mathematics and resources; check there is sufficient information and obtain any that is missing; use 'trial and improvement' methods. • use oral, written or visual forms to record and present findings. • make and test generalisations and simple hypotheses; define and reason in simple contexts with some precision.
Attainment Target 2	• count, read, write and order numbers to at least 10; know that the size of a set is given by the last number in the count. • understand the conservation of number.	• read, write and order numbers to at least 100; use the knowledge that the tens-digit indicates the number of tens. • understand the meaning of 'a half' and 'a quarter'.	• read, write and order numbers to at least 1000; use the knowledge that the position of a digit indicates its value. • use decimal notation as the conventional way of recording in money. • appreciate the meaning of negative whole numbers in familiar contexts.	• read, write and order whole numbers. • understand the effect of multiplying a whole number by 10 or 100. • use, with understanding, decimal notation to two decimal places in the context of measurement. • recognise and understand simple everyday fractions. • recognise and understand simple percentages. • understand and use the relationship between place values in whole numbers.	• use index notation to express powers of whole numbers. • use unitary ratios.	• read, write and order decimals; appreciate the relationship between place values. • understand and use equivalence of fractions and of ratios; relate these to decimals and percentages.
Attainment Target 3	• add or subtract, using objects where the numbers involved are no greater then 10.	• know and use addition and subtraction facts up to 10. • compare two numbers to find the difference. • solve whole number problems involving addition and subtraction, including money.	• know and use addition and subtraction number facts to 20 (including zero). • solve problems involving multiplication or division of whole numbers or money, using a calculator where necessary. • know and use multiplication facts up to 5 x 5, and all those in 2, 5 and 10 multiplication tables.	• know multiplication facts up to 10 x 10 and use them in multiplication and division problems. • (using whole numbers) add or subtract mentally two 2-digit numbers; add mentally several single-digit numbers; without a calculator add and subtract two 3-digit numbers, multiply a 2-digit number by a single-digit number and divide a 2-digit number by a single-digit number. • solve addition or subtraction problems using numbers with no more than two decimal places; solve multiplication or division problems starting with whole numbers.	• (using whole numbers) understand and use non-calculator methods by which a 3-digit number is multiplied by a 2-digit number and a 3-digit number is divided by a 2-digit number. • calculate fractions and percentages of quantities using a calculator where necessary. • multiply and divide mentally single-digit multiples of powers of 10 with whole number answers. • use negative numbers in context.	• work out fractional and percentage changes and related calculations. • calculate using ratios in a variety of situations. • convert fractions to decimals and percentages and find one number as a percentage of another.
Attainment Target 4	• give a sensible estimate of a small number of objects (up to 10).	• make a sensible estimate of a number of objects up to 20.	• recognise that the first digit is the most important in indicating the size of a number, and approximate to the nearest 10 or 100. • understand 'remainders' given the context of calculation; and know whether to round up or down.	• make use of estimation and approximation to check the validity of addition and subtraction calculations. • read a calculator display to the nearest whole number. • know how to interpret results on a calculator which have rounding errors.	• use and refine 'trial and improvement' methods. • approximate using a specified number of significant figures or decimal places.	• make use of estimation and approximation to check that the results of multiplication and division problems involving whole numbers are of the right order.
Attainment Target 5	• copy, continue and devise repeating patterns represented by objects/apparatus or one-digit numbers.	• explore and use the patterns in addition and subtraction facts to 10. • distinguish between odd and even numbers.	• explain number patterns and predict subsequent numbers where appropriate. • find number patterns and equivalent forms of 2-digit numbers and use these to perform mental calculations. • recognise whole numbers which are exactly divisible by 2, 5 and 10.	• apply strategies, such as doubling and halving, to explore properties of numbers, including equivalence of fractions. • generalise, mainly in words, patterns which arise in various situations.	• understand and use terms such as prime, square, cube, square root, cube root, multiples and factors. • recognise patterns in numbers through spatial arrangements. • follow simple sets of instructions to generate sequences.	• determine possible rules for generating a sequence. • use spreadsheets or other computer facilities to explore number patterns.
Attainment Target 6		• understand the use of a symbol to stand for an unknown number.	• deal with inputs to and outputs from simple function machines.	• understand and use simple formulae or equations expressed in words. • recognise that multiplication and division are inverse operations and use this to check calculations.	• understand and use simple formulae or equations expressed in symbolic form. • express a simple function symbolically.	• solve linear equations. • solve simple polynomial equations by 'trial and improvement' methods.
Attainment Target 7				• know the conventions of the co-ordinate representation of points; work with co-ordinates in the first quadrant.	• understand and use co-ordinates in all four quadrants.	• use and plot Cartesian coordinates to represent simple function mappings.

Attainment Target 8	• compare and order objects without measuring, and use appropriate language.	• use non-standard measures in length, area, volume, capacity, 'weight' and time to compare objects and recognise the need to use standard units. • know how to use coins in simple contexts. • know the most commonly used units in length, capacity, 'weight' and time, and what they are used for.	• use a wider range of metric units. • choose and use appropriate units and instruments in a variety of situations, interpreting numbers on a range of measuring instruments. • make estimates based on familiar units.	• understand the relationship between units. • find areas by counting squares, and volumes by counting cubes, using whole numbers. • make sensible estimates of a range of measures in relation to everyday objects or events.	• understand the notion of scale in maps and drawings. • use imperial units still in daily use and know their rough metric equivalents. • convert from one metric unit to another. • measure and draw angles to the nearest degree.	• understand and use compound measures. • recognise that measurement is approximate and choose the degree of accuracy appropriate for a particular purpose.
Attainment Target 9	• use materials provided for a task. • talk about own work and ask questions. • make predictions based on experience.	• select the materials and the mathematics to use for a task. • describe current work, record findings and check results. • ask and respond to the question: 'What would happen if . . .?'	• select the materials and the mathematics to use for a task; check results and consider whether they are sensible. • explain work being done and record findings systematically. • make and test predictions.	• select the materials and the mathematics to use for a task; plan work methodically. • record findings and present them in oral, written or visual form as appropriate. • use examples to test statements or definitions.	• select the materials and the mathematics to use for a task; check there is sufficient information; work methodically and review progress. • interpret mathematical information presented in oral, written or visual form.	• design a task and select appropriate mathematics and resources; check there is sufficient information and obtain any that is missing; use 'trial and improvement' methods. • use oral, written or visual forms to record and present findings. • make and test generalisations and simple hypotheses; define and reason in simple contexts with some precision.
Attainment Target 10	• sort 3-D and 2-D shapes. • build with 3-D solid shapes and draw 2-D shapes and describe them.	• recognise squares, rectangles, circles, triangles, hexagons, pentagons, cubes, rectangular boxes (cuboids), cylinders, spheres, and describe them. • recognise right-angled corners in 2-D and 3-D shapes.	• sort 2-D and 3-D shapes in different ways and give reasons for each method of sorting.	• understand and use language associated with angle. • construct simple 2-D and 3-D shapes from given information and know associated language.	• understand congruence of simple shapes. • explain and use angle properties associated with intersecting and parallel lines and triangles, and know associated language.	• know and use angle properties and symmetry properties of quadrilaterals and other polygons. • recognise and use common 2-D representation of 3-D objects. • use computers to generate and transform 2-D shapes. • classify and define types of quadrilaterals.
Attainment Target 11	• state a position using prepositions such as: on, inside, above, under, behind, next to, etc. • give and understand instructions for moving along a line.	• understand the notion of angle. • give and understand instructions for turning through right-angles. • recognise different types of movement: straight movement (translation); turning movement (rotation); flip movement (reflection).	• recognise the (reflective) symmetry in a variety of shapes in 2 and 3 dimensions. • understand eight points of the compass; use clockwise and anti-clockwise appropriately.	• specify location by means of co-ordinates (in first quadrant) and by means of angle and distance. • recognise rotational symmetry.	• identify the symmetries of various shapes. • use networks to solve problems. • specify location by means of co-ordinates in the four quadrants.	• understand and use bearings to define direction. • reflect simple shapes in a mirror line. • enlarge a shape by a whole number scale factor. • devise instructions for a computer to produce desired shapes and paths.
Attainment Target 12	• select criteria for sorting a set of objects and apply consistently.	• choose criteria to sort and classify objects; record results of observations or outcomes of events. • help to design a data collection sheet and use it to record a set of data leading to a frequency table.	• extract specific pieces of information from tables and lists. • enter and access information in a simple database.	• specify an issue for which data are needed; collect, group and order discrete data using tallying methods with suitable equal class intervals and create a frequency table for grouped data. • understand, calculate and use the mean and range of a set of data. • interrogate data in a computer database.	• design and use an observation sheet to collect data; collate and analyse results. • collect, group and order continuous data using equal class intervals and create a frequency table for grouped data. • insert and interrogate data in a computer database and draw conclusions.	• specify an issue for which data are needed; design and use an appropriate observation sheet to collect data; collate and analyse results. • design and use a questionnaire to survey opinion (taking account of bias); collate and analyse results.
Attainment Target 13	• record with real objects or drawings and comment about the result. • create simple mapping diagrams showing relationships; read and interpret them.	• construct, read and interpret block graphs and frequency tables. • use diagrams to represent the result of classifying using two different criteria.	• construct and interpret bar charts. • create and interpret graphs (pictograms) where the symbol represents a group of units.	• create a decision tree-diagram with questions to sort and identify a collection of objects. • construct, read and interpret a bar-line graph for a discrete variable (where the length of the bar-line represents the frequency). • construct and interpret a line graph and know that the intermediate values may or may not have a meaning. • construct and interpret a frequency diagram choosing suitable class intervals covering the range for a discrete variable.	• construct and interpret a pie chart from a collection of data with a few variables; interpret pie charts already presented in journals or newspapers. • construct and interpret conversion graphs. • construct and interpret a frequency diagram choosing suitable class intervals covering the range for a continuous variable.	• create scatter graphs for discrete and continuous variables and have a basic understanding of correlation. • construct, describe and interpret information through two-way tables. • construct and interpret work diagrams which represent relationships or connections.
Attainment Target 14	• recognise possible outcomes of simple random events.	• recognise that there is a degree of uncertainty about the outcome of some events and other events are certain or impossible.	• place events in order of 'likelihood' and use appropriate words to identify the chance. • understand and use the idea of 'evens' and say whether events are more or less likely than this. • distinguish between 'fair' and 'unfair'.	• understand and use the probability scale from 0 to 1. • give and justify subjective estimates of probabilities in a range of events. • list all the possible outcomes of an event.	• know that, when repeating the same experiment, different outcomes may result. • distinguish between estimates of probabilities based on statistical evidence and those based on the assumptions of symmetry. • know that if each of *n* events is assumed to be equally likely the probability of one occuring is 1/*n*.	• identify all the outcomes when dealing with two combined events which are independent using diagrammatic, tabular or other forms. • know that if there are several possible outcomes of an event (exhaustive and mutually exclusive), the total of these probabilities is 1 and that the probability of something happening is 1 minus the probability of it not happening.

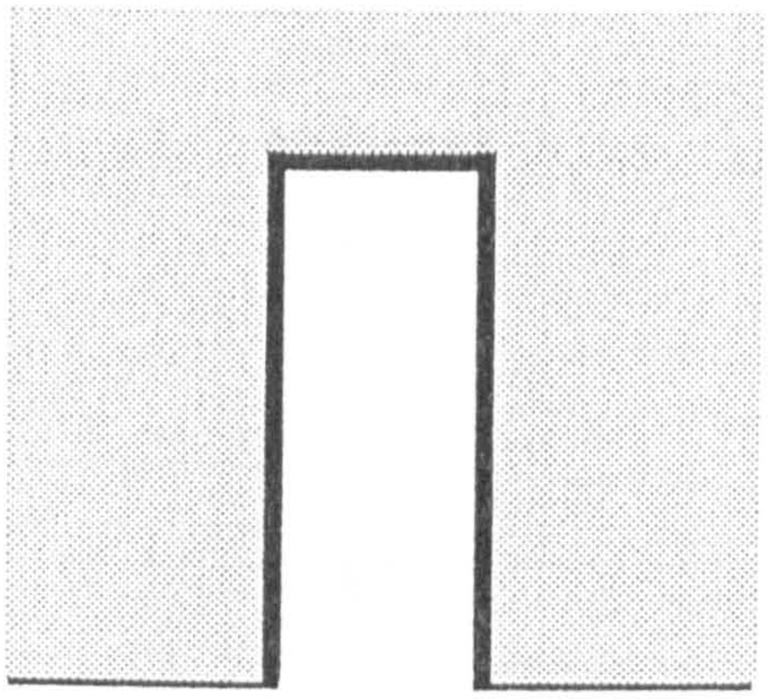
APPENDIX 5 (continued).

SD1 (1)

AT8 5a

Will these objects go through the doorway?

	Wardrobe	Chest of drawers
Height	210 cm	75 cm
Depth	50 cm	60 cm
Width	110 cm	90 cm



Scale: 1 cm to 50 cm

SD1 (2)

AT8 5a

This notice board is
230 cm by 130 cm.

Bill Poster has 5 notices of various sizes. Their dimensions are shown in the table.

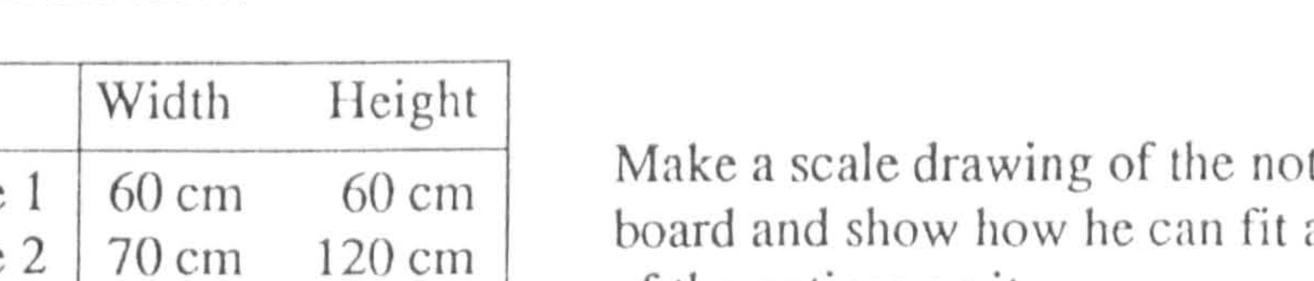
	Width	Height
Notice 1	60 cm	60 cm
Notice 2	70 cm	120 cm
Notice 3	70 cm	75 cm
Notice 4	70 cm	65 cm
Notice 5	80 cm	40 cm

Make a scale drawing of the notice board and show how he can fit all of the notices on it.

Use a scale of 1 cm to 10 cm.

SD1 (3)

AT8 5a

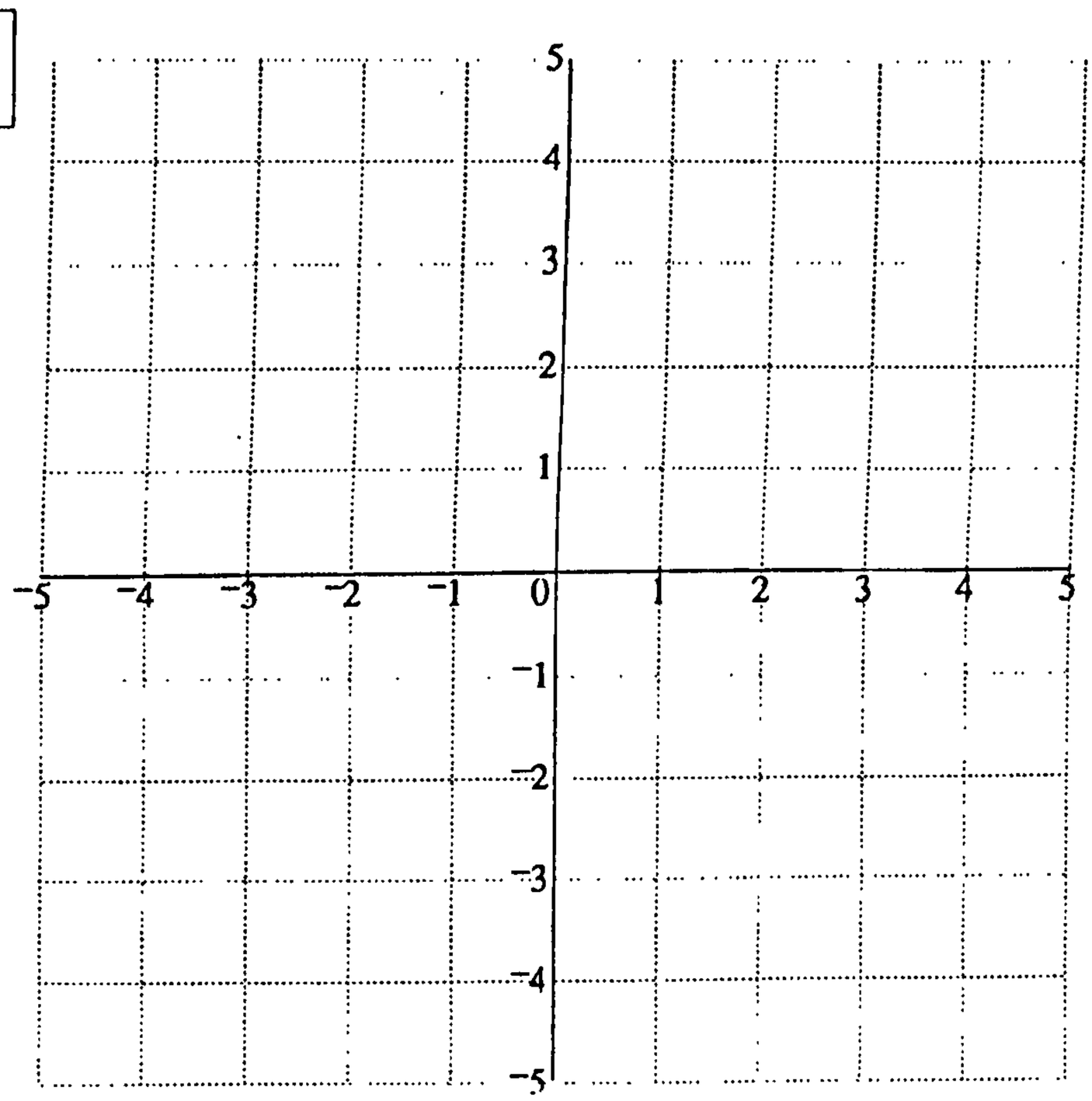


What do you think about the scale chosen?

226

This activity requires the pupil to use tracing paper and move it over the grid.

C2:e (3)
AT7 6a



Draw a straight line on tracing paper.
Place it on the grid so that all the points on the line fit:

(a) the rule $x = y$

(b) the rule $x - 2 = y$

C2:e (4)
AT7 6a

(a) Say which is the odd one out and why:

(0, 1) (-3,-5) (2, 5) (10, 21) (-1, -1)
(9, 12) (100, 201)

Write the rule for the other points using x, y language.

(b) Write down three rules for straight lines which pass through (2, 3).

(c) Find a rule which fits all of these points:

(1, 3) (3, 7) (5, 11) (6, 13)

FD1 (4)
AT2 4c
AT8 4a

This pole is 125 cm long.

This could also be written as 1 m 25 cm or 1·25 m.

Copy and complete this table writing the other lengths in different ways.

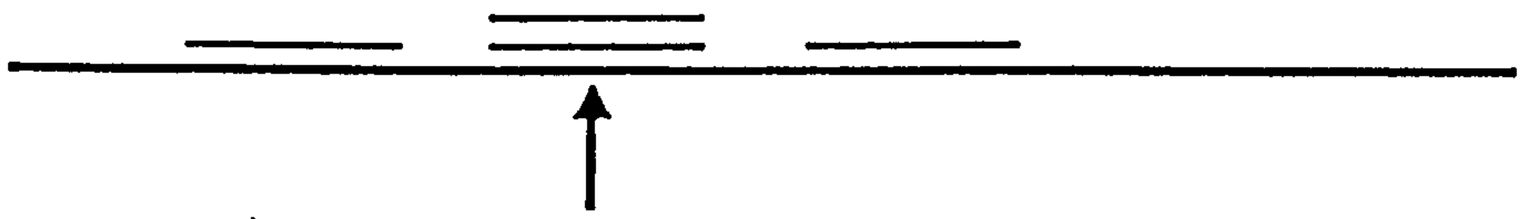
Length in m and cm	1 m 25 cm	3 m 80 cm			4 m 6 cm	
Length in cm	125 cm			65 cm		308 cm
Length in metres	1·25 m		2·5 m			

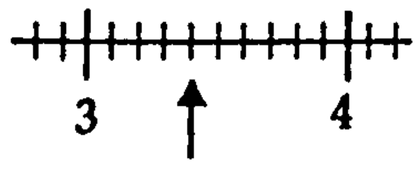
This is an activity for one or two pupils.
Instructions should be provided in written or oral form.
Discuss the arrangement with the pupil. (Aim to assess the reading of decimals e.g. not nought point ten for 0·10 and the understanding of place value.)

FD1 (5)
AT2 6a

Decimal cards

Arrange these cards in piles, from lowest to highest in value (lowest on the left, highest on the right).



Stack cards which have the same value
(e.g. 3·4 and  should be stacked together).

Introduction.

I hope you can offer assistance in the validation process I need to conduct with regard to some INSET materials I have prepared. The materials are to be used to assess the degree of reliability and validity of 'professional judgement' applied to teacher assessment under the national curriculum in key stage 3.

In order to analyse this aspect of teacher assessments it has been necessary to choose a selection of statements of attainment which reflect balance and breadth within the curriculum and to produce some test materials relating to these for use with pupils in key stage 3. Your task will be to 'judge' the materials in terms of a series of criteria - this will determine their appropriateness for the assessment of pupils.

Included here are three documents.

A - Test specifications. This details each SoA with its NC example and immediately below this are a set of specifications relating to the rules by which test items can be constructed to assess the above mentioned SoA. The specifications provide the necessary framework on which the test items are based and subsequently analysed. The specifications have 4 parts:

(1) this is a description of the SoA produced by CATS, the KS 3 pilot group - in theory it should make the SoA easier to interpret.
(2) this is a sample item(s) which may or may not be included in the test items.
(3) this outlines a series of criteria to delimit the type of questions to be asked; the aim is to ask the most generalisable type of question rather than a broad cross-section or range of questions - the latter approach hinders any fruitful analysis.
(4) this outlines what restrictions are or may be imposed in terms of the response of a pupil to the questions; i.e. which are acceptable responses and which are not.

B - Test Items. For each SoA and its accompanying test specification there are a series of test items. These test items are designed to be part of a test which would take place at, say, the end of term or half-term. Each item is denoted by a bracketed lower case letter e.g. (c). It is important to realise that at times whole questions may form a test item but, at other times, parts of questions may equally well form a test item. Any full or part question given a bracketed lower case letter is a 'test item'.

C - Summary Grid. There are two mains facets to Criterion Referenced assessment - (which is, in theory, what teacher assessment should emulate) - these are congruence and proficiency.

For the purposes of the materials presented here there are two congruence issues to consider. You will need to judge each test item's congruence with both the accompanying statement of attainment and the corresponding test specifications. Put more simply:

(i) do the test items 'fit' the SoA?
(ii) do the test items 'fit' the test specifications?

It should be noted that a yes to (i) does not necessarily mean a yes to (ii) and vica versa, although that is desirable.

Proficiency refers to the number or proportion of appropriate test items you consider a child would need to 'get right' on a particular topic for that child to be deemed proficient in that topic. In national curriculum terms the acknowledgement of proficiency is reflected in the awarding of the appropriate SoA.

Please follow these instructions very carefully.

(i) Read 'quickly' statement 1 and the accompanying test specifications for that statement.
(ii) Now look at the test items for statement 1.
(iii) Notice each item is denoted by a bracketed lower case letter.
(iv) Look at the summary grid.
(v) In the column labelled 1 fill in the appropriate gaps as prescribed by the criteria at the left hand side of the column - all entries will be yes/no or numerical. Notice some criteria may require you to make brief notes on the test item pages these should be done as instructed.
(vi) Any problems, then you may find the Explanatory Notes at the back of the summary grid useful.
(vii) Repeat this process for the other 9 Statements and corresponding test items.
(viii) On completion of all 10 statements the entire contents should be placed in the S.A.E and posted back to me. I would be grateful for their receipt by Wednesday 23rd OCT, at the very latest.

Thank you in anticipation,

Les Atkinson,

Document A - Test Specifications.

Statement (3.2a-K) 1. know and use addition and subtraction facts up to 10

Example: Know that if 6 pencils are taken from a box of 10, there will be 4 left.

Test Specifications:

(1) pupils should be able to add & subtract, mentally, numbers up to 10 and use this to help solve problems

(2) If 6 pencils are taken from a box of 10, how many will be left? If John has 3 sweets and Clare has 2 sweets, how many do they have altogether.

(3) (i) all questions should involve numbers up to a maximum of 10, with an adequate range used
(ii) items should depict practical situations i.e. ages, sweets, money
(iii) questions should be written and not just numerical in format i.e. 2 + 5 = is NOT acceptable represented in equal proportions
(iv) equal numbers of addition and subtraction questions should be present

(4) (i) correct answers only will be acceptable; addition and subtraction questions will be assessed seperately

---

Statement (3.3a-K) 2. know and use addittion and subtraction facts up to 20 (including zero)

Example: State that the date of the next Friday after Friday 8 May must be 15 May.

Test Specifications:

(1) pupils should be able to add & subtract, mentally, numbers up to 20 and use this to help solve problems

(2) Miklos has a piece of rope 14m long, he uses 5m to make a swing. How many metres does he have left? Jane is 4 years older than her 12 year old sister. How old is Jane?

(3) (i) all questions should involve numbers up to a maximum of 20, with an adequate range used
(ii) items should depict practical situations i.e. ages, sweets, money etc
(iii) questions should be written and not just numerical in format i.e. 12 + 5 = is NOT acceptable
(iv) equal numbers of addition and subtraction questions should be present

(4) (i) correct answers only will be acceptable; addition and subtraction questions will be assessed seperately

---

Statement 3. solve simple polynomial equations by 'trial and improvement' methods
(6.6b-S)
Example: Solve equations such as $x^2 = 5$ and $x^3 = 20$ using a calculator

---

Test Specifications:

(1) pupils should be able to solve equations like $x^3 = 21$ & $a^2 + 2 = 5$ by 'trial and improvement'

(2) Solve $x^2 = 5$ and $x^3 = 20$ using a calculator by using 'trial and improvement' methods. Solutions should be given to 2 decimal places.

(3) (i) pupils should be instructed not to use the square root or cube root keys on a calculator
(ii) the accuracy of the answer needs to be stated to 2 decimal places for each equation
(iii) numerical values in the equations should be restricted to integer values to 50 or less and indices to 2 and 3
(iv) an adequate coverage should be made of the numerical range for both index values of 2 & 3

(4) (i) solutions need to be stated to the specified accuracy
(ii) solutions must be accompanied by correct method
(iii) correct method allows for a single error within each calculation, i.e. one mistake per equation is acceptable but the appropriate solution must be commensurate with this error

---

Statement 4. use and plot Cartesian coordinates to represent
(7.6a-S) simple function mappings
Example: $x \dashrightarrow x + 1$ (or $y = x + 1$)
$x \dashrightarrow x^2$ (or $y = x^2$)

Test Secifications:

(1) pupils should be able to draw the graph of a simple function

(2) Plot the graphs of $y = x + 1$ and $x \dashrightarrow x^2$.

(3) (i) questions should not require the pupil to draw and label axes
(ii) questions should require the plotting of points for linear graphs of the form $y = mx + c$ ($m>0$ $c+/-$) and of simple quadratic graphs i.e. $y = x^2 + c$ ($c<6$)
(iii) axes should be 4 quadrant format with a scale of 1cm to 1unit
(iv) m & c values to be restricted therby allowing the function to be adequately represented on standard 2mm graph paper

(4) (i) pupils should plot points exactly when integer coordinates are involved and to +/- 1mm when fractional or decimal
(ii) linear graphs should have at least 2 points and quadratics at least 9 points e.g. $y = x^2$ with x values from -5 to +5 with integer incriments
(iii) one error per 9 po[illegible]ed is acceptable

Statement 5. understand the relationship between units
(8.4a-U)

Example: Use two units such as millilitres and litres to measure the capacity of the same jug.

Test Specifications:

(1) pupils should be able to see the connection between different units

(2) 3 litres = ...millilitres

(3) (i) all questions should involve measurements which will be of a familiar size to the pupil
(ii) items should cover the basic units of length, mass & capacity
(iii) conversions of milli/kilo and kilo/milli should be represented(centi/milli and centi/kilo and vica versa for length only)
(iv) conversion factors should not exceed 1000

(4) (i) correct answers only will be acceptable

Statement 6. understand the notion of scale in maps and
(8.5a-U) drawings

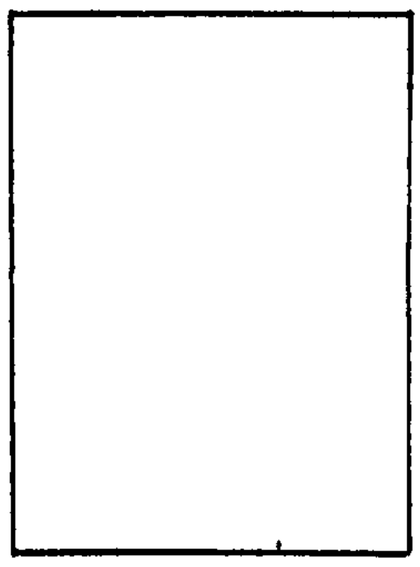
Example: Draw a plan of your classroom using a scale of 1cm to 1m.

Test Specifications:

(1) pupils should be able to use a scale on a map or drawing and explain what it means

(2) The diagram shows a scale drawing of a lawn. The scale is 1cm to 2m. How long and wide is the lawn?

A 1m border was made by removing lawn from the edges. Show this border on your diagram.



(3) (i) pupils should be provided with a partially complete scale drawing; consisting of a simple geometical shape i.e. a quadrilateral
(i) questions should provide the pupil with a scale in the form 1cm to Xunit format, where X is either 2, 5 or 10
(iii) Dimensions of the actual (full size) subject should be asked for
(iv) completion of the shape should be required; this should take the form of the construction of a border or of a simple extension of the shape in one of its dimensions

(4) (i) correct answers ony will be acceptable for the numeric portions of the item
(ii) construction of the border or extension should be to +/- 1mm in terms of its size and location

---

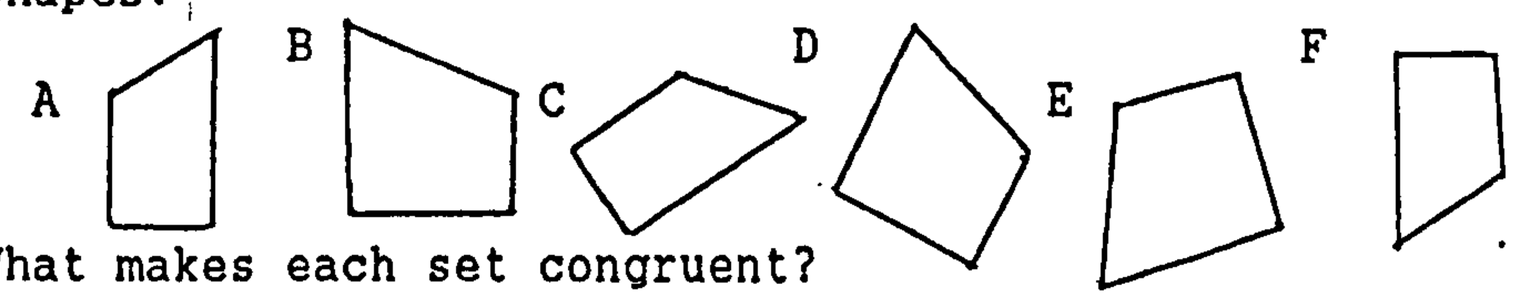
Statement 7. understand congruence of simple shapes
(10.5a-U)

Example: Group together congruent shapes from a range of shapes.

Test Specifications:

(1) pupils should be able to pick out things which are exactly the same size & shape and explain why they are the same, by describing their angles and the lengths of their sides

(2) Group together the following into sets of congruent shapes.



What makes each set congruent?

(3) (i) the shapes represented should be quadrilaterals but not squares or rectangles
(ii) relative orientations should not be at 90°, but multiples of this are permited
(iii) questions should not require pupils to draw or construct shapes
(iv) angles or lengths should not be given but incongruance should be apparent by significant length and angular differences

(4) (i) allow one mistake per congruent set, i.e. one shape misplaced or not chosen within that set; all answers need to be accompanied by a reason to be correct

---

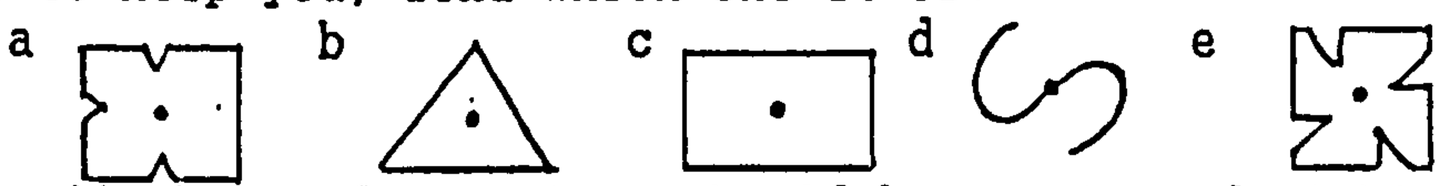
Statement 8. recognise rotational symmetry
(11.4b-S)

Example: Turn shapes using tracing paper.

Test Specifications:

(1) pupils should be able to recognise if shapes can be turned around to fit onto themselves

(2) When given a quarter turn about the dot, only one of these shapes will fit onto itself. Using a piece of tracing paper to help you, find which one it is.



(3) (i) all shapes for an item should be constucted within a unit cell i.e. a square, equilateral triangle or circle
(ii) there should be some representation of symmetric, asymmetric and non-symmetric shapes.
(iii) only quarter turn-symmetry should be included
(iv) pupils should not be required to draw or construct any shapes
(v) the use of tracing paper should be encouraged whenever possible
(vi) one of the distractors should posses no symmetry and at least one but not more than two others should have point symmetry; the shapes should have symmetry order 3 or 5

not cont. ¼ turn-

(4) (i) correct answers only will be acceptable

Statement 9. recognise that there is a degree of uncertainity
(14.2a-S) about the outcome of some events and other events are certain or impossible

Example: Recognise that it is:

certain	that 'it will get dark tonight'
impossible	that 'I will be 20 tomorrow'
uncertain	whether 'it will rain tomorrow'

Test Specifications:

(1) pupils should be able to think of some things that will definitely happen, definitely not happen and that `may or may not' happen

(2) Say if these are impossible, certain or uncertain.

"it will get dark tonight" = _____
"you will be 20 tomorrow" = _____
"it will rain tomorrow" = _____

(3) (i) questions should include at least one of each attribute but no more than two
(ii) pupils should be provided with events and required to indicate which single attribute they satisfy
(iii) events should be everyday and familiar to pupils and clearly within a specific attribute domain
(iv) numeric questions should not be used i.e. problems associated with dice or spinners etc.

(4) (i) correct answers only will be acceptable

Statement 10. understand and use the idea of `evens' and say
(14.3b-U) whether events are more or less likely than this

Example: Recognise that if a die is thrown there is an equal chance of an odd or even number, but the chance of getting a particular number (say 5), is less than an even chance.

Test Specifications:

(1) pupils should be able to give examples of things that have an `evens', better than `even' and a worse than `even' chance of happening

(2) When a die is thrown say if these outcomes are `evens', a `more' than evens or a `less' than evens chance of happening.

"a score of 2 or more" = _____
"an odd number" = _____
"a score of 5" = _____

(3) (i) questions should include at least one of each attribute but no more than two
(ii) pupils should be provided with events and required to indicate which single attribute they satisfy
(iii) events should be everyday and familiar to pupils and clearly within a specific attribute domain
(iv) numeric questions should be used, but with problems restricted to dice, coins or a pack of cards; events chosen should be less than 0.25, higher than 0.75 or 0.5 exactly in terms of probabilities

(4) (i) correct answers only will be acceptable

Document B - Test Items.

Test items for Statement 1.

Work out the following in your head:

(a) John has 3 sweets and Clare has 2 sweets. How many sweets do they have altogether?	H+	S	T
(b) Ian has 9 marbles to start with and wins 2 in a game. How many does he have now?			
(c) Mr Gupta has 1 daughter and 3 sons. How many children does he have?	H+	S	T
(d) Sarah has 4 pence and her Grandma gives her 5 pence. How much does Sarah have in total?	H+	S	T
(e) 6 pencils are taken from a box of 10. How many are left?	H-	S	T
(f) Verity has 9 comics and gives 3 of these to Tom. How many has she left?	H-	S	T
(g) I have 6 pence in my pocket. One coin is a 5 pence piece. What is the other coin?	H+	S	T
(h) Paul is 3 years younger than his 10 year old sister. How old is Paul?	H-	S	T

are answers >10 with the scope?

answer > 10?

Test items for Statement 2.

Work out the following in your head:

(a) John has 3 sweets and Clare has 2 sweets. How many sweets do they have altogether?	H+	S	T
(b) Ian has 9 marbles to start with and wins 2 in a game. How many does he have now?	H+	S	T
(c) Mr Gupta has 1 daughter and 3 sons. How many children does he have?	H+	S	T
(d) Sarah has 4 pence and her Grandma gives her 5 pence. How much does Sarah have in total?	H+	S	T
(e) 6 pencils are taken from a box of 10. How many are left?	H-	S	T
(f) Verity has 9 comics and gives 3 of these to Tom. How many has she left?	H-	S	T
(g) I have 6 pence in my pocket. One coin is a 5 pence piece. What is the other coin?	H-	S	T
(h) Paul is 3 years younger than his 10 year old sister. How old is Paul?	H-	S	T

These are not within "spirit" of NC statement but acceptable interpretations of the statement.

Test items for Statement 3.

Without using a squre root key, or cube root key, on your calculator, use a 'trial and improvement' method to solve these equations. Your solutions need to be correct to 2 decimal places.

H1 S

(a) $x^2 = 7$ H1 S (b) $x^2 = 15$ H1 S (c) $x^2 = 20$ H1 S (d) $x^2 = 33$ H1 S (e) $x^2 = 47$

H1 type

H2 S T

(a) $x^3 = 5$ H2 S T (b) $x^3 = 17$ H2 S T (c) $x^3 = 23$ H2 S T (d) $x^3 = 38$ H2 S T (e) $x^3 = 48$

H2 type

$x^2 + 0 =$ ↑ is zero not of $x + a = b$

Test items for Statement 4.

(i) On the axes provided plot the following graphs:

(a) $y = x$ H1 S T (b) $y = x+2$ H1 S T (c) $y = 2x-3$ H1 S T (d) $y = -x$ S (e) $y = -x+1$ S

SOA => SI
SPEL => h

(ii) On the axes provided plot the following graphs:

(a) $y = x^2$ H2 S T (b) $y = x^2+2$ H2 S T (c) $y = x^2-1$ H2 S T (d) $y = x^2-3$ H2 S T (e) $y = x^2+4$ H2 S T

Specification mg not familiar

APPENDIX 9 (continued)

Test items for Statement 5.

Fill in the blanks:

(a) 3l = ... ml (b) 70000ml = ... l

(c) 6000g = ... kg

(d) 5kg = ... g (e) 3000mg = ... g

(f) 9g = ... mg

(g) 4000m = ... km (h) 2km = ... m

(i) 2000mm = ... m (j) 5m = ... mm

(k) 2cm = ... mm (l) 30mm = ... cm

(m) 9m = ... cm (n) 400cm = ... m

H1 ÷ 1000 5/

H2 x 1000 5/1

H3 x 10 1/1

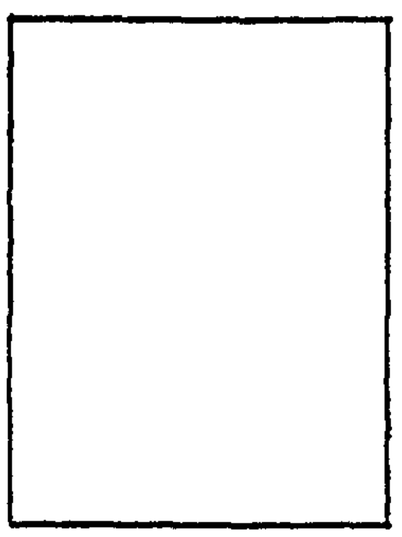
H4 ÷ 10 1/1

H5 x 100 1/1

H6 ÷ 100 1/1

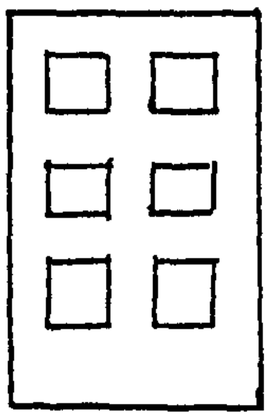
Test items for Statement 6.

(a) The diagram shows a scale drawing of a lawn. The scale is 1cm to 2m. How long and wide is the lawn?



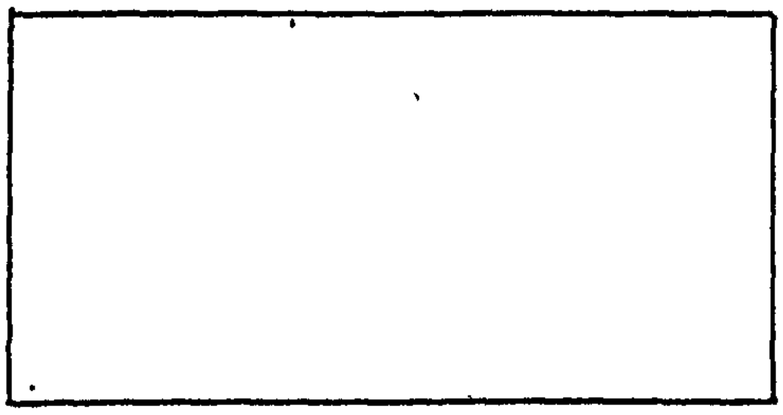
H S A 1m border was made by removing lawn from the edges. Show this border on your diagram.

(b) The diagram shows a building. The scale is 1cm to 10m. How tall and wide is the building?



H S An extra two floors are built which make the building 20m higher. Show this on your diagram.

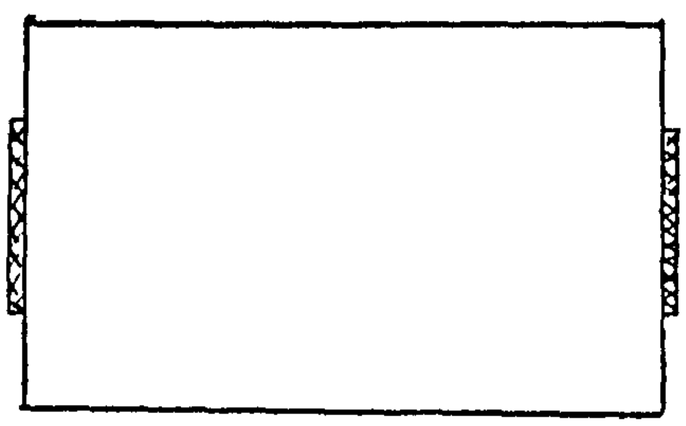
(c) The shop window shown is drawn to a scale of 1cm to 1m. Write down the actual length and width of the window.



Not 1 to 2,5,10

H: fractional ideas Show on your diagram, how the window could be divided into three equal parts.

(d) The 5-a-side football pitch, shown, is drawn with a scale of 1cm to 5m. How long and wide is the pitch?



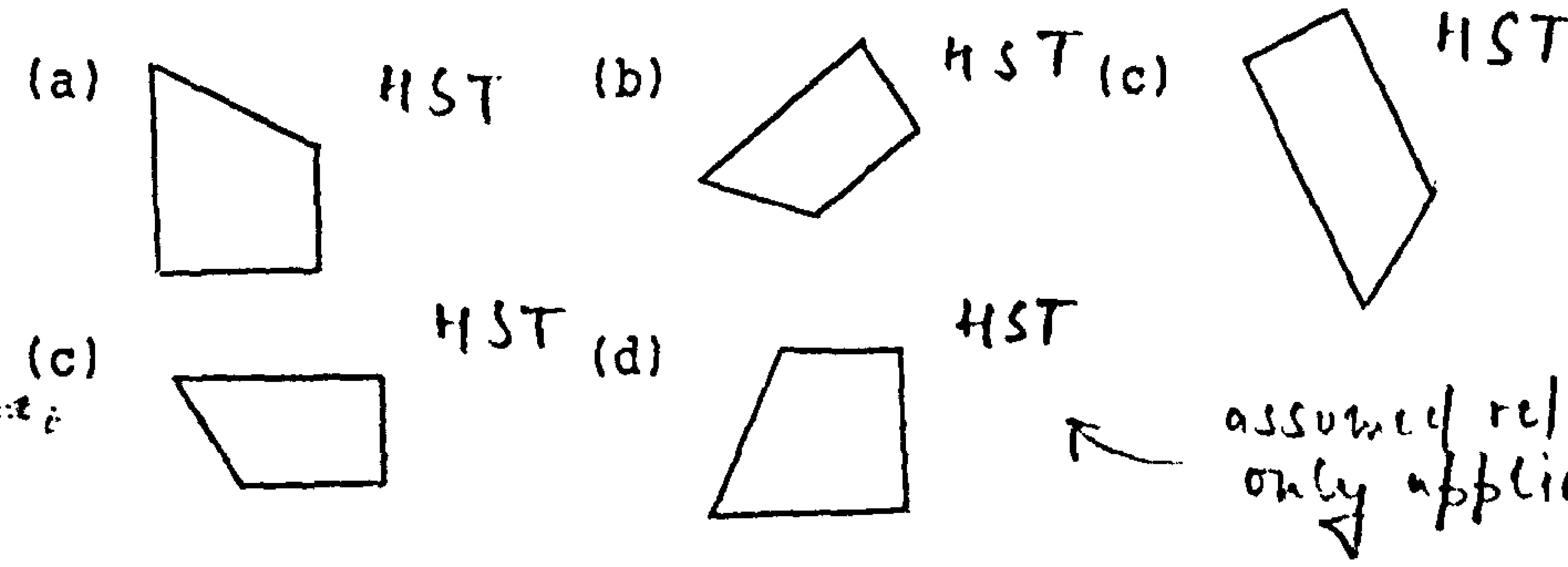
HST

ambiguous wording Draw on the diagram the 5m goal lines, i.e. lines across the pitch 5m from each goal; and the half way line.

Test items for Statement 7.

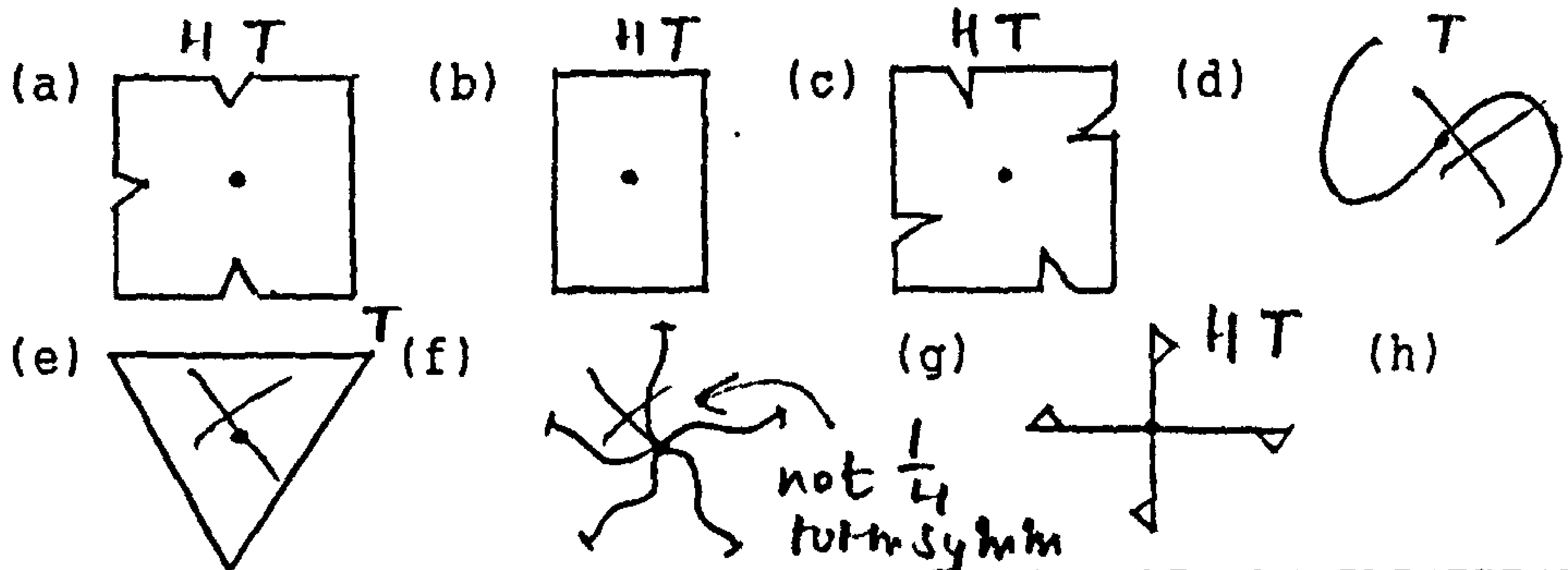
Look at this quadrilateral - 

Which of the following are congruent with the quadrilateral shown above and which are not? Give brief reasons for your answers.



Test items for Statement 8.

Which of these shapes will fit onto itself given a quarter turn. (hint: you may find a piece of tracing paper useful)



H (d)(e)(f) 1/4 turn hard to establish

S items do not seem to test recognition of rotational - but ability to follow instr.

Test items for Statement 9.

Say if the following are impossible, certain or uncertain:

HST (a) 'it will get dark tonight' c ______

HST (b) 'you will be 20 tomorrow' i ______

HST (c) 'it will rain tomorrow' u ______

HST (d) 'Tuesday will follow Monday' c ______

HS (e) 'score a 6 with the first throw of a dice' u

HST (f) 'a river will run uphill' i ______

& numeric / dice

Test items for Statement 10.

even, better, worse

Which of the above best describes the chance that:

HST (a) 'when you roll a dice you will score 2 or higher' ______ b

HST (b) 'you will get a 'tail' when you toss a coin' ______ e

HST (c) 'an odd number on a dice' ______ e — ideas of answer

HST (d) 'if you cut a pack of 52 cards you will get an ace' ______ w

HST (e) 'if you pick a day at random it will be a P(weekday): eek day' ______ b

HST (f) 'a score of 5 on a dice' ______ w

Document C - Summary Grid.

Summary Grid.

criteria		1	2	3	4	5	6	7	8
Homogeneity of Test Items: What proportion of the test items are doing basically the same assessment job: give as a fraction e.g. 5/6 etc. NB. see explanatory notes.		H+ 4/8 H- 3/8 H 7/8	H+ 4/8 H- 4/8 H 8/8	H1 5/10 H2 5/10 ::	H1 3/5 H2 5/5 H 8/10	SEE TEST ITEMS	3/4	5/5	4/7
Statement of Attainment: What proportion of the test items are congruent with the National Curriculum statement of attainment: give as a fraction. N.B. see explanatory notes.		7/8	8/8	10/10	10/10	0/14 fraction	3/4	5/5	0/7
Test Secifications: What proportion of the test items are congruent with the test specifications: give as a fraction. N.B. see explanatory notes.		7/8	8/8	5/10	8/10	14/14 but do not cover full spec	3/4	5/5	6/7
Proficiency Ratings: What fraction of the test items would you require the pupil to correctly respond to before crediting the statement. - e.g. 2/4 subtractions etc. NB. for questions with more than one distinct type could you rate each portion and also give an overall rating as and when applicable.	TYPE I	add 4/8	add 4/8	H1 4/5	H1 3/5	1000 ÷ × 8			
	TYPE II	sub 4/8	sub 4/8	H2 4/5	3/5	100 ÷ × 2			
	TYPE III					10 ÷ × 2			
	OVER ALL	8/8	8/8	8/10	6/10	12/14	3/4	5/5	6/7
Minimum Number Of Items: Could you please indicate the minimum number of items you would deem adequate to give an appropriate opportunity for the demonstration of each statement of attainment. e.g. for 3.2a you may decide 4 items are needed for addition and 4 items for the subtraction parts to give the pupil enough scope to show what they can do.		3 (+) 3 (-)	3 (+) 3 (-)	(H1) 3 (H2) 3	(H1) 2 (H2) 2	(H1/H2) 3 (H3/H4) 3 (H5/H6) 3	3.	5	5

Any problems see the explanatory notes!

Explanatory Notes:

Homogeneity of Test Items: Which items are 'doing the same assessment job' i.e. which items would you group together as being of the same sort or similar, you can give this as a fraction e.g. 3 out of 4 say. Indicate on the test item page which ones are a part of the homogenous group by a capital H next to the item - for those which aren't homogenous could you indicate, very briefly, the reason why not next to the item.

Statement of Attainment: Which items fit the National Curriculum statement of attainment, in your judgement. Indicate this with a capital S, for those that don't fit could you, very briefly, say why not next to the item.

Test Specifications: Which items fit the test specifications, in your judgement. Indicate this with a capital T, for those that don't fit could you, very briefly, say why not next to the item.

Proficiency Ratings: Some questions are straight forward and you can give a simple fractional answer to the proportion of items you would expect a pupil to respond correctly to award the SoA. Others, however, are not so straight forward. For example the measurement question has three different types of items. (distance/length/mass) To avoid pre-judging I have provided a 3 part section for you to comment on the individual types of question involved along with an overall section. Please feel free to comment on this on the grid if you have any views.

Minimum Number of Items: You may consider there are too few items to allow a pupil to adequately demonstrate proficiency, or lack of it, in the corresponding statement of attainment. On the other hand you may feel there are too many items. Whichever the view, could you give the minimum number of test items you would judge to be necessary for the purposes of assessing a pupil on a particular statement of attainment.

Any major problems then please call me at home on 0377 - 241367 or at school 0723 - 582174.

Any Comments please give below.

mel

Questionnaire - Grantley/York - 11/91.

As a consequence of the preceding discussion could you please complete the following details:

Part I - Personal Details.

(a) D.O.B...4/11/63.....

(b) Number of years teaching Mathematics 5

(c) Is this your principal subject YES ✓ NO ___ If NO what is?

(d) Previous experience of Criterion-Referenced assessment other than with the National Curriculum. YES ___ NO ✓ If YES, please specify details, briefly.

(e) Since when have you been involved with Teacher Assessments - please give a date, e.g. Sept' 89. If you haye not had experience of this write NONE.

Sept 89.

Part II - Congruence.

Do you consider that the majority (75% or more) of the SoAs within the National Curriculum (old version) provide sufficient detail to allow you to identify appropriate assessment materials. YES ✓ NO ___ If NO briefly state why not.

Part III - Proficiency Ratings.

Given the following number of items linked to a particular SoA, what proportion or fraction of these would you expect the pupil to correctly respond to for the statement to be given:

	Work completed in the classroom					
Number of items	2	3	4	5	6	10
Fraction or Proportion	2/2	3/3	4/4	4/5	5/6	8/10

	Work completed under test conditions					
	2	3	4	5	6	10
	2/2	3/3	3/4	4/5	4/6	7/10

Part IV - Short Term Recall.

Does Short Term Recall have an effect on the assessment of a pupils' work?
YES ✓ NO ___

If YES, what do you consider to be the minimum period of time required before short term recall has no longer a significant effect?

Time period as a number of days, weeks, months, etc. 3weeks.

Any Other Comments:

APPENDIX 12.

PULL OUT INFORMATION

'Piece-A'

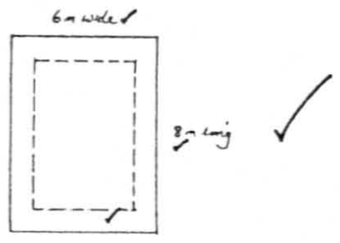
Work out the following in your head:

1. John has 3 sweets and Clare has 2 sweets. How many sweets do they have altogether? 5 ✓
2. Ian has 9 marbles to start with and wins 1 in a game. How many does he have now? 11 ✗
3. Sarah has 4 pence and her Grandma gives her 5 pence. How much does Sarah have in total? 8 ✗
4. 6 pencils are taken from a box of 10. How many are left? 5 ✗
5. Ann has 9 comics and gives 3 of these to Tom. How many has she left? 6 ✓
6. Paul is 3 years younger than his 10 year old sister. How old is Paul? 7 ✓

'Piece-O'

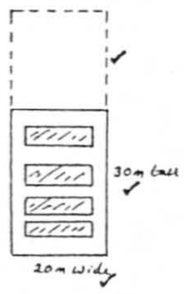
1. The diagram shows a scale drawing of a lawn. The scale is 1cm to 2m. How long and wide is the lawn?

   A 1m border was made by removing lawn from the edges. Show this border on your diagram.



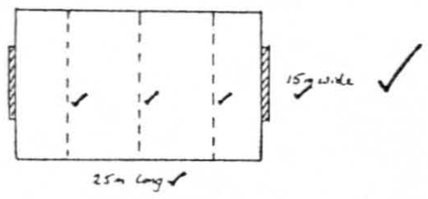
2. The diagram shows the side view of a building. The scale is 1cm to 10m. How tall and wide is the building?

   An extra two floors are built which make the building 20m higher. Show this on your diagram.



3. The 5-a-side football pitch, shown, is drawn with a scale of 1cm to 5m. How long and wide is the pitch?

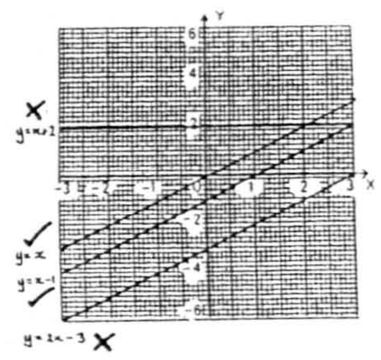
   Draw on the diagram the 5m goal lines, i.e. lines across the pitch 5m from each goal and the half way line.



'Piece-T'

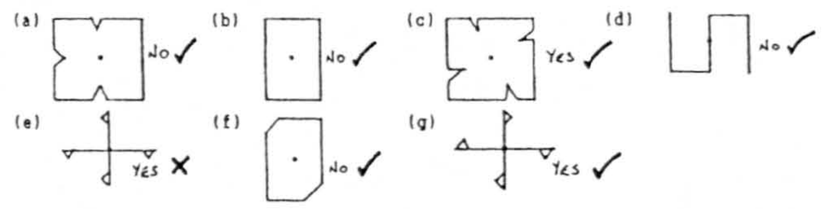
On the pair of axes provided plot the following graphs:

1. y = x 2. y = x+2 3. y = 2x-3 4. y = x-1



'Piece-K'

Which of these shapes will fit onto itself given a quarter turn. (hint: you may find a piece of tracing paper useful)



'Piece-H'

('EVENS') ('BETTER THAN EVENS') ('WORSE THAN EVENS')

Which of the above best describes the chance that:

(a) 'you will get a 'tail' when you toss a coin' evens ✓

(b) 'if you cut a pack of 52 cards you will get an ace' worse ✓

(c) 'if you roll a dice you will get a score of 5' worse ✓

Page 1 - please read this before you answer the questions below

Listed below are 5 statements of attainment (SoAs). Each SoA is paired with a piece of pupils' work - the same upper case letter is used to indicate the pairings. You need to look at the SoA with its particular piece of work and decide whether or not the pupil has 'attained' the statement based on what is shown. Tick YES or NO in the space provided. You should spend no longer than 1 minute on each question. There are a further 5 pairings of SoAs and pieces of pupils' work, overleaf. So, it is expected you will spend 10 minutes in total to complete this questionnaire.

Additional Information.

All the work shown on pages 2 and 4 was completed by pupils under formal test conditions as part of an end of term assessment.

The 10 pieces were completed by 10 different pupils, taught by 10 different teachers, across years 7 and 8.

The 10 SoAs and pupils' work have been arbitrarily allocated an upper case letter to allow them to be readily identified.

	'Statement Attained'	
Statement of Attainment - A 'know and use addition and subtraction facts up to 10'	YES ___	NO ___
Statement of Attainment - T 'use and plot Cartesian coordinates to represent simple function mappings'	YES ___	NO ___
Statement of attainment - O 'understand the notion of scale in maps and drawings'	YES ___	NO ___
Statement of Attainment - K 'recognise rotational symmetry'	YES ___	NO ___
Statement of attainment - H 'understand and use the idea of 'evens' and say whether events are more or less likely than this'	YES ___	NO ___

PTO..

For identification purposes could you put you D.O.B here please.....

Notes:

The pupils' work has been reduced to half size for practical purposes

Remember you should spend approximately 1 minute per question this will ensure the authenticity of the exercise - Thank you

PULL OUT INFORMATION

'Piece-E'

Work out the following in your head:

1. Ian is 9 years older than his 1 year old sister. How old is Ian? 10 ✓
2. Bob has 4 stamps to start with and is given 2 more. How many does he have now? 6 ✓
3. Christine has 5 pence and she finds another 3 pence. How much does she have in total? 8 ✓
4. I have 10 pence in my pocket and take out 6 pence. What is left? 4 pence ✓
5. Jenny is 5 years younger than her 6 year old brother. How old is Jenny? 1 ✓
6. 3 pens are taken from a box of 9. How many pens are still left in the box? 6 ✓

'Piece-S'

Without using the cube root key, on your calculator, use a 'trial and improvement' method to solve these equations. Your solutions need to be correct to 2 d.p.'s

1. $x^3 = 17$

$2.5^3 = 15.625$
$2.6^3 = 17.576$
$2.55^3 = 16.58$
Solution is 2.55 ✗

2. $x^3 = 23$

$2.8^3 = 21.952$
$2.9^3 = 24.389$
$2.85^3 = 23.149$
$2.84^3 = 22.906$
$2.83^3 = 22.665$
Solution is 2.84 ✓

3. $x^3 = 38$

$3.5^3 = 42.875$
$3.4^3 = 39.304$
$3.35^3 = 37.595$
$3.36^3 = 37.933$
$3.37^3 = 38.273$
Solution is 3.36 ✓

4. $x^3 = 48$

$3.8^3 = 54.872$
$3.7^3 = 50.653$
$3.65^3 = 48.627$
$3.64^3 = 48.228$
Solution is 3.64 ✗

'Piece-J'

Fill in the blanks:

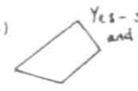
1. 3litres = .3000. millilitres ✓
2. 5kg = .5000. g ✓
3. 7km = .7000. m ✗
4. 2cm = ..20... mm ✓
5. 9m = .9000 cm ✗

'Piece-Q'

Look at this quadrilateral – 

Which of the following are congruent with the quadrilateral shown above and which are not? Give brief reasons for your answers.

(a)  No - not the same sides and angles. ✓

(b)  Yes - same angles and sides. ✓

(c)  No - not same sides. ✓

(c)  Yes - same angles and sides. ✓

(d)  No - not same sides or angles. ✓

'Piece-D'

Say if the following are impossible, certain or uncertain:

(a) 'it will get dark tonight' uncertain ✗
(b) 'you will be 20 tomorrow' impossible ✓
(c) 'it will rain tomorrow' uncertain ✓
(d) 'Tuesday will follow Monday' certain ✓
(e) 'I will come top in this maths test' uncertain ✓
(f) 'a river will run uphill' impossible ✓

Page 3 – have you read the instructions overleaf? if yes, carry on!

	'Statement Attained'
Statement of Attainment - E 'know and use addition and subtraction facts up to 20 (including zero)'	YES ___ NO ___
Statement of Attainment - S 'solve simple polynomial equations by "trial and improvement" methods'	YES ___ NO ___
Statement of attainment - J 'understand the relationship between units'	YES ___ NO ___
Statement of Attainment - Q 'understand the congruence of simple shapes'	YES ___ NO ___
Statement of attainment - D 'recognise that there is a degree of uncertainty about the outcome of some events and other events are certain or impossible'	YES ___ NO ___

For identification purposes could you put your D.O.B here please .....

Notes:

The pupils' work has been reduced to half size for practical purposes

Remember you should spend approximately 1 minute per question this will ensure the authenticity of the exercise – Thank you

APPENDIX 13.

PULL OUT INFORMATION

'Piece-A'

Work out the following in your head:

1. John has 3 sweets and Clare has 2 sweets. How many sweets do they have altogether? 5 ✓
2. Ian has 9 marbles to start with and wins 1 in a game. How many does he have now? 11 ✗
3. Sarah has 4 pence and her Grandma gives her 5 pence. How much does Sarah have in total? 8 ✗
4. 6 pencils are taken from a box of 10. How many are left? 5 ✗
5. Ann has 9 comics and gives 3 of these to Tom. How many has she left? 6 ✓
6. Paul is 3 years younger than his 10 year old sister. How old is Paul? 7 ✓

'Piece-J'

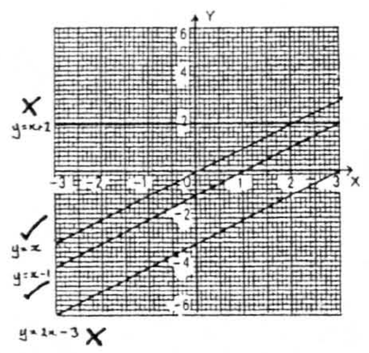
Fill in the blanks:

1. 3litres = 3000 millilitres ✓
2. 5kg = 5000 g ✓
3. 7km = 7000 m ✗
4. 2cm = 20 mm ✓
5. 9m = 9000 cm ✗

'Piece-T'

On the pair of axes provided plot the following graphs:

1. y = x
2. y = x+2
3. y = 2x-3
4. y = x-1

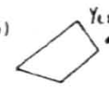


'Piece-Q'

Look at this quadrilateral -

Which of the following are congruent with the quadrilateral shown above and which are not? Give brief reasons for your answers.

(a)  No - not the same sides and angles. ✓

(b)  Yes - same angles and sides. ✓

(c)  No - not same sides. ✓

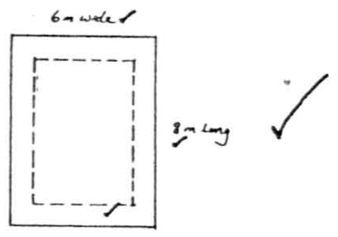
(d)  No - not same sides or angles. ✓

(e)  Yes - same angles and sides. ✓

'Piece-O'

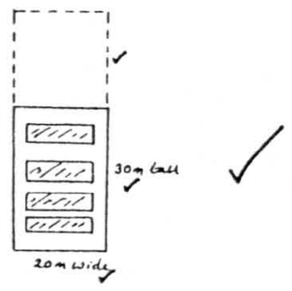
1. The diagram shows a scale drawing of a lawn. The scale is 1cm to 2m. How long and wide is the lawn?

   A 1m border was made by removing lawn from the edges. Show this border on your diagram.



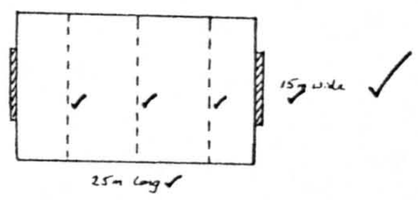
2. The diagram shows the side view of a building. The scale is 1cm to 10m. How tall and wide is the building?

   An extra two floors are built which make the building 20m higher. Show this on your diagram.



3. The 5-a-side football pitch, shown, is drawn with a scale of 1cm to 5m. How long and wide is the pitch?

   Draw on the diagram the 5m goal lines, i.e. lines across the pitch 5m from each goal and the half way line.



Page 1 - please read this before you answer the questions below

Listed below are 5 statements of attainment (SoAs). Each SoA is paired with a piece of pupils' work - the same upper case letter is used to indicate the pairings. You need to look at the SoA with its particular piece of work and decide *whether or not the pupil has 'attained' the statement based on what is shown.* Tick YES or NO in the space provided. You should spend no longer than 1 minute on each question. There are a further 5 pairings of SoAs and pieces of pupils' work, overleaf. So, it is expected you will spend 10 minutes in total to complete this questionnaire.

Additional Information.

All the work shown on pages 2 and 4 was completed by pupils under formal test conditions as part of an end of term assessment.

The 10 pieces were completed by 10 different pupils, taught by 10 different teachers, across years 7 and 8.

The 10 SoAs and pupils' work have been arbitrarily allocated an upper case letter to allow them to be readily identified.

	'Statement Attained'
Statement of Attainment - A 'know and use addition and subtraction facts up to 10'	YES ___ NO ___
Statement of Attainment - J 'understand the relationship between units'	YES ___ NO ___
Statement of attainment - T 'use and plot Cartesian coordinates to represent simple function mappings'	YES ___ NO ___
Statement of Attainment - Q 'understand the congruence of simple shapes'	YES ___ NO ___
Statement of attainment - O 'understand the notion of scale in maps and drawings'	YES ___ NO ___

PTO..

Notes:

The pupils' work has been reduced to half size for practical purposes

Remember you should spend approximately 1 minute per question this will ensure the authenticity of the exercise - Thank you

PULL OUT INFORMATION

'Piece-D'

Say if the following are impossible, certain or uncertain:

(a) 'it will get dark tonight' — uncertain ✗

(b) 'you will be 20 tomorrow' — impossible ✓

(c) 'it will rain tomorrow' — uncertain ✓

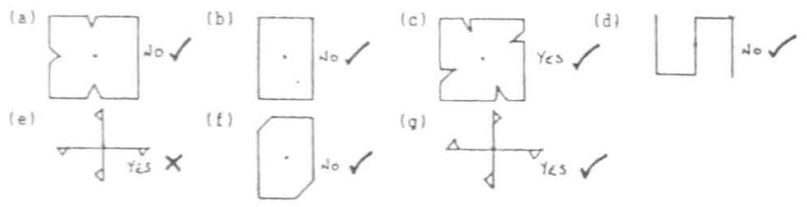
(d) 'Tuesday will follow Monday' — certain ✓

(e) 'I will come top in this maths test' — uncertain ✓

(f) 'a river will run uphill' — impossible ✓

'Piece-K'

Which of these shapes will fit onto itself given a quarter turn. (hint: you may find a piece of tracing paper useful)



'Piece-E'

Work out the following in your head:

1. Ian is 9 years older than his 1 year old sister. How old is Ian? 10 ✓
2. Rob has 4 stamps to start with and is given 2 more. How many does he have now? 6 ✓
3. Christine has 5 pence and she finds another 3 pence. How much does she have in total? 8 ✓
4. I have 10 pence in my pocket and take out 6 pence. What is left? 4 pence ✓
5. Jenny is 5 years younger than her 6 year old brother. How old is Jenny? 1 ✓
6. 3 pens are taken from a box of 9. How many pens are still left in the box? 6 ✓

'Piece-H'

('EVENS') ('BETTER THAN EVENS') ('WORSE THAN EVENS')

Which of the above best describes the chance that:

(a) 'you will get a 'tail' when you toss a coin' — evens ✓

(b) 'if you cut a pack of 52 cards you will get an ace' — worse ✓

(c) 'if you roll a dice you will get a score of 5' — worse ✓

'Piece-S'

Without using the cube root key, on your calculator, use a 'trial and improvement' method to solve these equations. Your solutions need to be correct to 2 d.p.'s

1. $x^3 = 17$

$2.5^3 = 15.625$

$2.6^3 = 17.576$

$2.55^3 = 16.58$

Solution is 2.55 ✗

2. $x^3 = 23$

$2.8^3 = 21.952$

$2.9^3 = 24.389$

$2.85^3 = 23.149$

$2.84^3 = 22.906$

$2.83^3 = 22.665$

Solution is 2.84 ✓

3. $x^3 = 38$

$3.5^3 = 42.875$

$3.4^3 = 39.304$

$3.35^3 = 37.595$

$3.36^3 = 37.933$

$3.37^3 = 38.273$

Solution is 3.36 ✓

4. $x^3 = 48$

$3.8^3 = 54.872$

$3.7^3 = 50.653$

$3.65^3 = 48.627$

$3.64^3 = 48.228$

Solution is 3.64 ✗

Page 3 - have you read the instructions overleaf? if yes, carry on!

	'Statement Attained'
Statement of Attainment - D 'recognise that there is a degree of uncertainty about the outcome of some events and other events are certain or impossible'	YES ___ NO ___
Statement of Attainment - K 'recognise rotational symmetry'	YES ___ NO ___
Statement of attainment - E 'know and use addition and subtraction facts up to 20 (including zero)'	YES ___ NO ___
Statement of Attainment - H 'understand and use the idea of 'evens' and say whether events are more or less likely than this'	YES ___ NO ___
Statement of attainment - S 'solve simple polynomial equations by "trial and improvement" methods'	YES ___ NO ___

For identification purposes could you put your D.O.B here please .....

Notes:

The pupils' work has been reduced to half size for practical purposes

Remember you should spend approximately 1 minute per question this will ensure the authenticity of the exercise - Thank you

APPENDIX 14.

PULL OUT 'FLYSHEET' INFORMATION

'Piece-A'

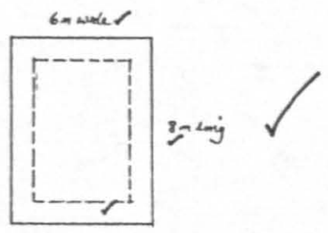
Work out the following in your head:

1. John has 3 sweets and Clare has 2 sweets. How many sweets do they have altogether? 5 ✓
2. Ian has 9 marbles to start with and wins 1 in a game. How many does he have now? 11 ✗
3. Sarah has 4 pence and her Grandma gives her 5 pence. How much does Sarah have in total? 8 ✗
4. 6 pencils are taken from a box of 10. How many are left? 5 ✗
5. Ann has 9 comics and gives 3 of these to Tom. How many has she left? 6 ✓
6. Paul is 3 years younger than his 10 year old sister. How old is Paul? 7 ✓

'Piece-O'

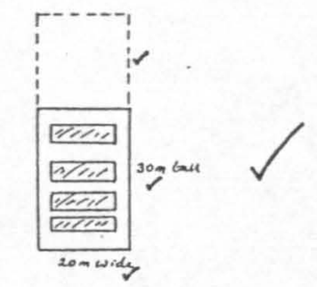
1. The diagram shows a scale drawing of a lawn. The scale is 1cm to 2m. How long and wide is the lawn?

   A 1m border was made by removing lawn from the edges. Show this border on your diagram.



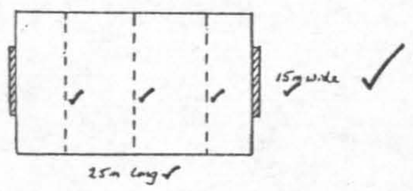
2. The diagram shows the side view of a building. The scale is 1cm to 10m. How tall and wide is the building?

   An extra two floors are built which make the building 20m higher. Show this on your diagram.



3. The 5-a-side football pitch, shown, is drawn with a scale of 1cm to 5m. How long and wide is the pitch?

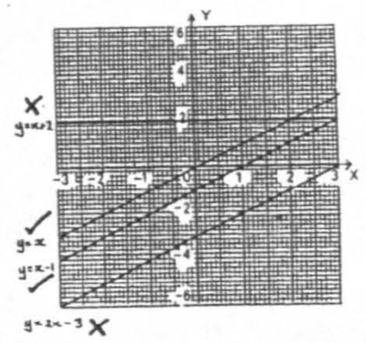
   Draw on the diagram the 5m goal lines, i.e. lines across the pitch 5m from each goal and the half way line.



'Piece-T'

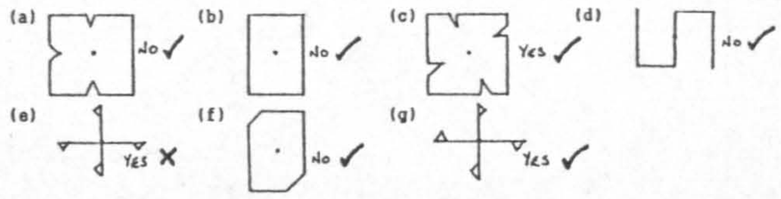
On the pair of axes provided plot the following graphs:

1. y = x 2. y = x+2 3. y = 2x-3 4. y = x-1



'Piece-K'

Which of these shapes will fit onto itself given a quarter turn. (hint: you may find a piece of tracing paper useful)



'Piece-H'

'EVENS' 'BETTER THAN EVENS' 'WORSE THAN EVENS'

Which of the above best describes the chance that:

(a) 'you will get a 'tail' when you toss a coin' evens ✓

(b) 'if you cut a pack of 52 cards you will get an ace' worse ✓

(c) 'if you roll a dice you will get a score of 5' worse ✓

	CON	THR	MAX
H	0	1	1
A	1	0	0
T	0	0	0
O	1	1	1
K	1	1	0

'Piece-E'

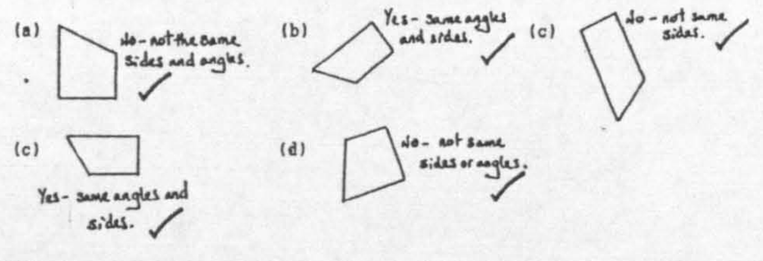
Work out the following in your head:

1. Ian is 9 years older than his 1 year old sister. How old is Ian? 10 ✓
2. Rob has 4 stamps to start with and is given 2 more. How many does he have now? 6 ✓
3. Christine has 5 pence and she finds another 3 pence. How much does she have in total? 8 ✓
4. I have 10 pence in my pocket and take out 6 pence. What is left? 4pence ✓
5. Jenny is 5 years younger than her 6 year old brother. How old is Jenny? 1 ✓
6. 3 pens are taken from a box of 9. How many pens are still left in the box? 6 ✓

'Piece-Q'

Look at this quadrilateral - 

Which of the following are congruent with the quadrilateral shown above and which are not? Give brief reasons for your answers.



	CON	THR	MAX
E	0	1	1
S	1	0	0
J	0	0	0
Q	1	1	1
D	1	1	0

'Piece-S'

Without using the cube root key, on your calculator, use a 'trial and improvement' method to solve these equations. Your solutions need to be correct to 2 d.p.'s

1. $x^3 = 17$

   $2.5^3 = 15.625$
   $2.6^3 = 17.576$
   $2.55^3 = 16.58$
   Solution is 2.55 ✗

2. $x^3 = 23$

   $2.8^3 = 21.952$
   $2.9^3 = 24.389$
   $2.85^3 = 23.149$
   $2.84^3 = 22.906$
   $2.83^3 = 22.665$
   Solution is 2.84 ✓

3. $x^3 = 38$

   $3.5^3 = 42.875$
   $3.4^3 = 39.304$
   $3.35^3 = 37.595$
   $3.36^3 = 37.933$
   $3.37^3 = 38.273$
   Solution is 3.36 ✓

4. $x^3 = 48$

   $3.8^3 = 54.872$
   $3.7^3 = 50.653$
   $3.65^3 = 48.627$
   $3.64^3 = 48.228$
   Solution is 3.64 ✗

'Piece-D'

Say if the following are impossible, certain or uncertain:

(a) 'it will get dark tonight' uncertain ✗

(b) 'you will be 20 tomorrow' impossible ✓

(c) 'it will rain tomorrow' uncertain ✓

(d) 'Tuesday will follow Monday' certain ✓

(e) 'I will come top in this maths test' uncertain ✓

(f) 'a river will run uphill' impossible ✓

'Piece-J'

Fill in the blanks:

1. 3litres = .3000. millilitres ✓
2. 5kg = .5000. g ✓
3. 7km = .7000. m ✗
4. 2cm = ..20... mm ✓
5. 9m = ..9000 cm ✗

APPENDIX 15.

PULL OUT INFORMATION

Page 1a - Instructions - please read this page carefully before you go any further.

I would be grateful if you could do the following:

1. Complete the personal details below.
2. Ensure you spend no more than 1 **MINUTE** answering each of the questions on pages 1b and 3. This is vital to provide for authentic and realistic assessment.
3. Look at page 1b, answering the questions by making reference to page 2.
4. Remove the paperclip, open the A3 sheet and read through page 3 answering the questions by making reference to page 4.
5. DO NOT make any assumptions about the pupils' work, other than it has been carefully marked - with correct answers indicated by a tick. You need to assess the work as it is shown.
6. It is essential you work individually on this exercise.
7. Give the completed scripts to your Head of Department for dispatch.
8. You will need to spend approximately 20mins in total - 10mins for the questions on pages 1b and 3 and 10mins for reading & completing your personal details.

I thank you in anticipation of your cooperation,

Les Atkinson.

Personal Details:

(a) Number of years teaching Mathematics _____

(b) Is this your principal subject Yes ____ No ____ If No, what is?

(c) Excluding National Curriculum assessment, have you any other experience of Criterion Referenced assessment e.g. NEA coursework, CPVE. Yes ___ No ___
If Yes, please give brief details.

(d) Could you indicate the date you started to make Teacher assessments for your National Curriculum groups. i.e. When did you start making assessments against Statement of attainments. Please give a date, e.g. Sept '89.

(e) D.O.B _____ this is optional but it would be useful.

Now turn to the back of this A3 sheet to page 1b - DON'T remove the paperclip yet.

Page 1b - have you read page 1a, if yes then please read this before you answer the questions below.

Listed below are 5 statements of attainment (SoAs). Each SoA is paired with a piece of pupils' work - the same upper case letter is used to indicate the pairings. You need to look at the SoA with its particular piece of work and decide whether or not the pupil has 'attained' the statement based on what is shown. Tick YES or NO in the space provided. You should spend no longer than 1 minute on each question. There are a further 5 pairings of SoAs and pieces of pupils' work, within this folded A3 sheet. So, it is expected you will spend 10 minutes in total to complete these 10 questions.

Additional Information.

All the work shown on page 2 was completed by pupils under formal test conditions as part of an end of term assessment; the work on page 4 formed part of a topic test which was given TWO weeks after the topic was completed, i.e. two weeks after any directly related teaching.

The 10 pieces were completed by 10 different pupils, taught by 10 different teachers, across years 7 and 8.

The 10 SoAs and pupils' work have been arbitrarily allocated an upper case letter to allow them to be readily identified.

N.B. - the pupils' work for the statements below can be found on page 2

'Statement Attained'

Statement of Attainment - E

'know and use addition and subtraction facts up to 20 (including zero)' YES ___ NO ___

Statement of Attainment - S

'solve simple polynomial equations by "trial and improvement" methods' YES ___ NO ___

Statement of attainment - J

'understand the relationship between units' YES ___ NO ___

Statement of Attainment - Q

'understand the congruence of simple shapes' YES ___ NO ___

Statement of attainment - D

'recognise that there is a degree of uncertainty about the outcome of some events and other events are certain or impossible' YES ___ NO ___

Now could you remove the paperclip & work through page 3 inside - thank you.

APPENDIX 15 (continued).

PULL OUT INFORMATION

Page 3a – Would you please read this carefully, you may find it useful!

When assessing criteria (as with the SoAs) there is a well defined two stage process which should be followed:

Stage 1. Congruence - does the work 'match' or 'fit' the SoA

Ask the following questions of the pupils' work:

(a) Is the SoA ADDRESSED by the work

(b) Is it at the appropriate level of DIFFICULTY - comparing the work with an e.g. is useful here

(c) is the work GENERALISABLE - for instance which is better for testing the tables a series of questions like 1x2 or a series like 8x7 - the latter would be considered to be more generalisable, i.e. success on the latter would probably mean success on most tables questions!

Stage 2. Proficiency - how many do you need to get right?

Consider these general proficiency ratings:

(a) GCSE exam boards expect between 58% - 66% for attainment of their target grades .

(b) The Key Stage 1 SAT expected 'all' or 'all but one' correct for the attainment of a statement.

(c) Mastery Learning usually demands 80% correct.

(d) The Chelsea Diagnostic testing team specified 2/3rds correct for the attainment of their criteria.

The above information may be a useful guide for you when you are looking at the SoAs and pupils' work on pages 3b & 4.

Summary:

Congruence and adequate proficiency are essential if a statement is to be attained. To avoid the difficulty of being overcritical when determing a piece of works' congruence or incongruence, it may be easier to classify work into the following groups:

Work which is clearly incongruent	-	classify as INCONGRUENT
Work which is clearly congruent	-	classify as CONGRUENT
Work which you are not sure of	-	classify as CONGRUENT

In essence you give the 'might be' work the benefit of the doubt!

Could you now complete page –3b making reference to page 4 when necessary.

Page 3b – have you read the information on page 3a, if yes carry on.

Additional information.

As mentioned overleaf, the work on page 2 was completed by pupils under formal test conditions as part of an end of term assessment; the work on page 4 formed part of a topic test which was given TWO weeks after the topic was completed, i.e. two weeks after any directly related teaching.
Each SoA given below is accompanied by its National Curriculum example.
For this exercise you will need to respond Yes or No to the Statement Attained question but also you will be given the opportunity to indicate your opinion on the aspects of 'congruence' and 'proficiency'; i.e. whether you think the work is congruent with the SoA, Yes/No; and if there is adequate proficiency, Yes/No.

N.B. – the pupils' work for the statements below can be found on page 4

	'Statement Attained'	Cong'	Prof'
Statement of Attainment - A 'know and use addition and subtraction facts up to 10' Example know that if 6 pencils are taken from a box of 10, there will be 4 left	Yes__ No__	Y__ N__	Y__ N__
Statement of Attainment - T 'use and plot Cartesian coordinates to represent simple function mappings' Example x --> x + 1 (or y = x + 1) x --> x^2 (or $y = x^2$)	Yes__ No__	Y__ N__	Y__ N__
Statement of attainment - O 'understand the notion of scale in maps and drawings' Example draw a plan of your classroom using a scale 1cm to 1m	Yes__ No__	Y__ N__	Y__ N__
Statement of Attainment - K 'recognise rotational symmetry' Example turn shapes using tracing paper	Yes__ No__	Y__ N__	Y__ N__
Statement of attainment - H 'understand and use the idea of 'evens' and say whether events are more or less likely than this' Example recognise that if a die is thrown there is an equal chance of an odd or even number, but the chance of getting a particular number (say 5) is less than an even chance	Yes__ No__	Y__ N__	Y__ N__

could you now hand in the completed questionnaire to your Head of Department and s/he will dispatch it – thank you

TABLE III

Average hours for each group of activities				
Activities	Secondary		Primary	
	Average hours	%	Average hours	%
Teaching	18.36	33.06	22.03	39.72
Preparing/planning for teaching	10.12	15.89	15.08	24.92
Administration teaching	8.73	12.35	6.97	8.77
Cover/Exam Invigilation	1.11	1.62	0.47	0.30
Pastoral/Disciplinary Activities	2.67	3.92	0.83	0.65
Duties/Supervising pupils	3.91	6.42	3.40	4.53
Parental Communication	2.59	2.64	2.66	2.37
General Administration	6.82	8.29	4.86	5.27
Inset	2.53	2.58	2.69	2.81
Break from work	2.53	4.06	2.73	4.53
Staff meetings/Discussions	5.15	8.17	3.62	5.68
Commuting	0.99	1.09	0.68	0.44

NON-CONTACT TIME

The average non-contact time for PRIMARY teachers was 34 minutes per WEEK i.e. approximately 7 minutes per DAY. Four-fifths of primary teachers had less than 12 minutes per day. The average non-contact time for SECONDARY teachers was 0.25 with two thirds having less than 0.2. This represents approximately 60 minutes per DAY.

DELEGATED TIME

The following figures were recorded in the section where teachers were asked to indicate how much of their time could have been delegated to a non teaching assistant.

Primary A total of 172 full time teachers completed this section.

The total amount of time was 871.25 hours, giving an average of 5.07 hours per teacher. If it is assumed that the other 14 primary teachers did not feel they could delegate any time and failed to record '0' then the average per week becomes 4.69 hours per teacher.

These figures applied to ALL full time primary teachers in Humberside represents between 1.37 (using 5.07) and 1.26 (using 4.69) additional full time NTAs per primary schools.

Assumptions made are an average of 10 teachers per primary school and a full time NTA works 37 hours per week.

Secondary A total of 110 full time teachers completed this section providing a total time of 345.75 hours. This is an average of 3.14 hours per week per teacher. When the average is calculated using all 121 full time teachers the figure obtained is 2.86 hours. When these figures are applied to ALL secondary teachers it represents between 5.1 (3.14) and 4.64 (2.86) NTAs per school.

Analyses of 'Activity Sampling' carried out during the Autumn Term 1990

TABLE I shows the distribution between Primary, Secondary, Nursery and Special Schools of the 316 FULL-TIME teachers who responded together with the breakdown of salary scales.

TABLE I

Primary		%	Secondary	%	Nursery	Special
NSS	96	51.6	37	30.6	2	1
+A	24	12.9	13	10.7	1	-
+B	33	17.7	18	14.9	3	-
+C	2	1.1	10	8.3	-	1
+D	-		24	19.8	-	-
+E	-		6	5.0	-	-
DH	19	10.2	9	7.4	-	1
H	12	6.5	4	3.3	-	-
Totals 186		100.00	121	100.00	6	3

'AVERAGE WORKING WEEK'

TABLE II gives the average number of hours worked in the particular week in each sector as well as a breakdown by salary scale. The maximum and minimum number of hours worked by individuals are also given. The astonishing figure of 112 hours was achieved by someone responsible for a weekend field trip!

TABLE II

Primary (186 FT)				Secondary (121FT)		
	Hours			Hours		
	Max	Min	Average	Max	Min	Average
All Teachers	81.0	39.75	55.46	112	34.75	55.08
Standard Scale	69.75	39.75	54.35	112	34.75	53.24
+ A	81.0	44.25	55.89	67.50	41.0	53.85
+ B	70.0	43.75	56.14	74.0	37.50	52.89
+ C				66.25	48.25	55.55
+ D				79.75	45.75	57.01
+ E				82.75	45.75	60.96
D. Head/Heads	72.75	47.25	58.50	83.0	42.50	57.90
Nursery (6)	69.25	55.00	59.29			
Special (3)	60.25	49.50	55.17			

HOW THE TIME WAS SPENT

It is virtually impossible to actually proportion time to discrete activities since many are carried out simultaneously by teachers but table III is an attempt to obtain some idea of the most likely distribution of time on the activities that were listed. Activities have been grouped together where they were of a 'similar' nature. The percentage is of the 'average' working week.

INSET - YORK P.M.

2 issues: (i) Congruence

(ii) Proficiency

(i) Congruence means does the work 'match' or 'fit' the SoA

- is the SoA addressed by the work
- is it at the appropriate level of difficulty
- is the work the most GENERALISABLE

e.g. testing the tables which is better

1x2 or 8x7?

NOTE: it is easier to determine non-congruence than congruence as with proof etc.

(ii) Proficiency means how many do you need to get right?

- Exam boards 58% -> 66% for GCSE target grades
- KS 1 SAT 'all' or 'all but one rule'
- Mastery Learning criteria 80%
- Chelsea Diagnostic Tests 2/3rds

NOTE: there are no simple answers to this one and in some respects it can be an arbitrary decision made in the first instance.

IMPORTANT - if congruence is not met then there is little point in considering proficiency.