VOLUME ONE

PROBLEM-SOLVING IN THE CONTEXT OF THE GENERAL CERTIFICATE OF EDUCATION ORDINARY LEVEL CHEMISTRY EXAMINATION

A thesis submitted for the degree of
Doctor of Philosophy of the
University of Keele

David Slimming

September 1984

ABSTRACT

The work described in this thesis concerns the problem-solving behaviour of students towards short-answer, non-mathematical items in Ordinary level chemistry examinations. The literature of problem-solving is selectively reviewed and an information processing model is advanced to provide a theoretical framework for the study. In the first phase of the work a protocol approach was used to generate detailed descriptions of students' problem-solving strategies and errors. It was found that problem-solving performance did not depend on IQ and that the strategies and errors of high and low achievers showed few differences. It was hypothesised that problem-solving behaviour was largely a reflection of the organisation of knowledge in a student's semantic memory.

In the second phase of the study an analysis-of-variance design was used to examine the influence, on students' problem-solving, of the level of information provided in examination items. It was found that performance tended to fall as levels of information rose, irrespective of whether the information was "relevant" or "irrelevant" to the problem to be solved. Limitations in the ways in which students appeared to perceive such information are discussed and the implications of these for problem-solving in realistic situations, where the information available may not be well matched to the demands of the task, are considered.

The second phase of the study also examined the hypothesis that differences in problem-solving behaviour might be attributable to features in semantic memory. Following a selective review of work on cognitive mapping, a word association technique was used to investigate students' semantic networks in selected topic areas, namely

electrochemistry, redox and atomic structure. Systematic differences, which could be related to details of students' problem-solving behaviour, were found between the maps of high and low achievers.

The educational implications of the findings and some suggestions for further research are discussed in the final chapter.

ACKNOWLEDGEMENTS

I gratefully acknowledge the thoughtful advice and kind encouragement which I have received from my supervisor, Professor Richard Kempa. His sure guidance has been invaluable.

I am also grateful to the British Council for the grant of a fellowship to enable me to undertake the study, and to numerous members of its staff who have provided help and support in various ways. I particularly wish to thank Mrs Marion Robinson, who typed the thesis, for her patience and skill.

Finally, I must acknowledge the contributions of my wife Pat, and my children, Sarah, Meyric and Amanda. They have willingly put up with considerable dislocation of our family life and have borne patiently with a husband and father who has sometimes been rather preoccupied! Without their cheerful and loving co-operation in many particular ways, I would not have been able to see the work through.

CONTENTS

CHAPTER 1

THE CONTEXT OF THE INVESTIGATION

The aim of this chapter is to explain the focus of interest of the investigation, to place it in the broad context of previous research and to draw attention to some of the factors which were considered in the early stages of planning the work. Many of the issues raised will, of course, be pursued in much greater depth in subsequent chapters. A secondary objective will be to provide a brief overview of the conduct of the investigation and of the organisation of this report.

1.1 Focus on examinations

The importance which our society places on examinations hardly needs to be stressed. The great majority of children regularly face internal school examinations in a range of subjects throughout their school careers. These examinations provide feedback to the teacher and the student regarding the apparent progress of learning, and the results contribute to decisions regarding the "setting" of students and the selection of subjects for study at subsequent stages. They also provide essential practice for the all important public examinations which most students will attempt at the end of compulsory schooling. Performance in these will often have a crucial bearing on a student's immediate and long-term prospects. Those obtaining high enough marks normally proceed to further academic training and further examinations. Yet, despite the way in which the educational system is dominated by examinations, relatively little is known of the cognitive processes associated with examination success. A single grade may be the only visible end-product of several years of study, but it is far from clear what such a grade indicates in terms of the cognitive capabilities of the individual. Recent proposals for the adoption of "grade-related criteria" in

Scotland (Weston, 1983), experiments with graded tests in parts of England and Wales (Harrison, 1982) and with criterion-referencing (Kempa and L'Odiaga, 1983) represent moves in this direction but are largely based on an analysis of performance on examination tasks rather than on empirical research into cognitive processes. The current investigation arose from a concern to understand more about how candidates actually tackle examination items in science, and thus to gain a better insight into the cognitive behaviours which such examinations reward.

In the case of school science, curriculum change in recent decades has placed considerable importance upon developing in pupils the ability to apply as well as to recall information. Modern examinations are routinely constructed so as to include predetermined proportions of items intended to reflect particular "levels" of intellectual demand. However, factor analysis studies have failed either to provide empirical validation for such levels or to generate useful alternatives (eg Lewis, 1961, 1964, 1969a, b; Chokotho, 1975; L'Odiaga, 1977). Massey (1982) has recently shown that even students obtaining the top A and B grades in the General Certificate of Education Ordinary Level chemistry examination show very limited mastery in terms of their ability to solve items demanding more than the recall of information. And it is still common experience that successful examinees may be unable to demonstrate the competencies expected of them outside the examination hall. So exactly what skills is the examination measuring? What, in the examination hall, are the cognitive events that determine success or failure? What strategies do students use and how do those of the more successful candidate differ from those of the less successful? What sorts of errors are made and how do such errors arise? Given the importance placed by science educators on the achievement of higher level curricular goals, and by society at large on examination grades, it would obviously be of value to know the answers to such questions. The answers may have implications not only for examining and the interpretation of examination marks but more widely for the way in which we teach. Cowan (1977) has argued persuasively that knowledge of the different ways in which students attempt problems can help us "to enhance the quality of their learning experience and subsequent performance". Similarly, more detailed knowledge of the ways in which errors arise may enable us to identify weaknesses in our teaching and contribute to the development of improved pedagogical methods.

An attempt to address some of the issues raised above provided the starting point for the present study. The General Certificate of Education Ordinary Level (GCE O-level) chemistry examination was selected as the particular context, on the one hand because of the background and interest of the author as a some-time teacher of chemistry, and on the other because of the importance of this level of examination at the end of compulsory schooling. Practical considerations relating to syllabuses and to numbers of accessible students influenced the selection of the GCE rather than the Certificate of Secondary Education examination.

1.2 The relationship of the present study to previous research

The present study, then, was concerned with the exploration of the problem-solving[1] processes of students in the formal assessment situation as represented by the GCE O-level chemistry examination.

---

1. The term "problem-solving" will be used throughout the study to refer to the cognitive behaviour of candidates as they attempt any examination item. Although this does not accord fully with classical definitions of problem-solving, some recent studies (Greeno, 1980b; Gabel and Sherwood, 1983) have advocated usage consistent with that adopted here. This issue is more fully discussed in the next chapter.

Both assessment and problem-solving have attracted some research interest; unfortunately they have generally been treated as separate issues and most studies are therefore of only limited relevance. Although widespread interest in assessment is reflected, for example, in the lengthy series of Schools Council Examination Bulletins, publications of this sort have been largely concerned with the development and evaluation of the objectives and techniques of examination. The Assessment of Performance Unit, set up within the Department of Education and Science in 1975 and charged with monitoring national levels of performance in a number of subject areas including science, has also been principally concerned with developing methods of assessment. In the course of this work the Unit has developed an interesting assessment framework based on "categories of science performance" (Harlen, Black and Johnson, 1981). However, like the categories customarily used in compiling science examinations, these are based upon an analysis of objectives and/or tasks and not on research into cognitive functioning. Research studies relating to assessment in science reported in Eggleston and Kerr (1968), and others referred to by Keohane (1977), show a similar emphasis. In recent years, the Assessment Group of the Royal Society of Chemistry has published collections of papers under the title "Research in Assessment" (Kempa, Hoare and Rutherford, 1975; Hoare, Kempa and Ongley, 1978; Ongley, Brockington and Sheridan, 1983). These include an account by Curragh and Savage (1975) of a study of the relationships between students' attainments in GCE O-level chemistry and their scores on tests of Raven's matrices and on tests of Piagetian formal operations tasks. Although some significant correlations were found between psychological measures and certain chemistry sub-scores (the derivation of which are not fully reported in the source referred to), no useful conclusions emerged. Otherwise the question of the cognitive processes of examination candidates is addressed only in relation to the **objectives** of assessment. By contrast the concern of the present study was to explore **actual** cognitive behaviours in the assessment situation.

Cognitive behaviour has, of course, been a central concern of research in the field of problem-solving. Although traditionally focusing on relatively content-free puzzles (that is to say problems which demand little or no knowledge of a specialised nature) there has recently been increasing interest in mathematics problems and in content-rich domains including science. Some of the early work relating to problem-solving in science has been criticised as shallow (Belanger, 1969) and much of it has tended to focus on the teaching and learning of a priori sets of problem-solving skills. There has, more recently, been some direct exploration of the problem-solving processes of science students, however this has tended to relate to the teaching/learning, rather than to the examination, context. The tasks employed have often contained a strong mathematical element and have tended to involve students in further and higher education rather than at school level. The data has generally been collected in the form of tape-recorded protocols obtained from the students as they tackle the problems. Much of the work has been of a descriptive/case-study nature and few principles of widely established generality have yet emerged. Nor, on account of differing problem types and differing contexts, is it immediately clear how such research relates to the concerns of the present study. It will be convenient, therefore, to postpone more detailed discussion of these issues until the next chapter.

1.3 A preliminary consideration of factors likely to influence examination performance

It seems reasonable to suppose that the kinds of processes used in solving problems will depend on the characteristics of both the problem and the problem-solver (Greeno, 1980). The factors considered below have therefore been categorised as task variables and student variables.

1.3.1 Task variables. Throughout the study, a clear distinction is maintained between different tasks on the one hand and different formulations of essentially the same task on the other. A problem-solving task is defined by the information given and demanded in the task statement. Clearly a student's problem-solving behaviour may vary between one task and another. It may also vary if the same task is presented in different ways.

Ashmore, Frazer and Casey (1979) have classified chemical problems (that is, different chemical tasks) on two dimensions, namely the nature of the solution required and the sources of information which must be employed. The former range from closed problems with a unique solution to open-ended problems such as the present study! Sources of information may be confined to the problem-statement or may call on memory, on authoritative sources such as books or ultimately on observation or experimentation as in fundamental research. Examination items, certainly in the context of written O-level chemistry papers, are closed or relatively so and demand the use of information from the problem-statement and memory only. The candidate's task is generally defined by fairly concise statements of the information given and the information to be generated. His attempt is constrained by the contents of his memory and by his ability to manipulate the available data.[1]

Examination items are generally classified in terms of both their chemical content and the level of educational objective, often based on Bloom's taxonomy (Bloom, 1956), which the item is intended to assess. Although it is **not** the purpose of the present study to evaluate the

1. Masculine forms of pronouns have occasionally been used throughout the thesis when referring to individuals who might be of either gender.

latter, it will obviously be important to ensure that tasks are selected for study in a systematic way and that they adequately represent an appropriate range in terms of likely intellectual demand as well as content. A further distinction can usefully be drawn between tasks which demand mathematical manipulation and those which do not. The difficulties which many students experience with items involving mathematics is well known to both chemistry teachers and researchers, and a numerical or mathematical factor has been identified in factor analysis studies of test materials in science (eg Lewis, 1964, 1969a; Slimming, 1971).

A particular task may be presented to the candidate in more than one way. Written O-level chemistry papers normally employ:

(i) objective items (normally multiple-choice and related types),

(ii) structured or "short-answer" items, and

(iii) essay-type items.

Clearly these may tend to elicit different cognitive behaviours in the context of essentially similar tasks (see also Kempa and L'Odiaga, 1983). For example multiple-choice items are likely to provide more cues and, indeed, miscues. They also offer opportunities for elimination and guessing in a way that other styles of item do not. Essay questions on the other hand would be expected to make heavier demands on a candidate's ability to select and combine relevant information as well as upon linguistic skills. Even within a given type of item the formulation of the task may be altered in more subtle ways: by changes in language and vocabulary; by presenting information symbolically or in tabular form rather than verbally and so on. Some research information is already available on the influence of language in examination items (Cassels and Johnstone, 1977, 1980, 1983) but there is obviously much scope for investigation into the influence of other aspects of task formulation.

1.3.2 Student variables. Many of a student's attributes might be expected to influence his problem-solving behaviour. Such attributes are considered below under two main headings: affective factors and cognitive factors. Cognitive style and gender are also discussed briefly as variables of potential interest.

(i) **Affective factors.** A number of positive attitudes have long been associated with successful problem-solving. Bloom and Broder (1950) have summarised these in a list which includes interest, self-confidence, concentration, application, perseverance and a careful and systematic approach. Poor problem-solvers tend to have the opposite characteristics and examination anxiety is widely held to be associated with poor performance. A recent finding that mathematics anxiety is negatively related to problem-solving scores on certain text-book chemistry questions is particularly relevant (Gabel and Sherwood, 1983). Although no systematic investigation into affective factors was undertaken during the present study, the need was recognised to make positive efforts to encourage the interest and to minimise the anxiety of the students participating.

(ii) **Cognitive factors.** Cognitive factors are obviously of central importance in any study of the problem-solving process. In particular the role of general intelligence (IQ) and of memory may be important and preliminary consideration is given to these below.

A number of studies have suggested a strong connection between measures of "intelligence" and achievement in science and in chemistry (eg Mallinson, 1969; Carlson, 1977). Lewis (1969a) found that performance in science was positively related to the general intellectual factor (g) and to two other factors, one of which contained elements of spatial (s), numerical (n), and verbal (v) abilities. It has been reported elsewhere that verbal ability is a good predictor of performance on science process tasks (Van Neie, 1972). On this basis the possible role of general intelligence was taken into consideration when planning the investigation.

The heavy reliance which examination candidates are obliged to place on memory has already been mentioned. Not only is memory the source of most of the factual information required in answering items, it is also the source of knowledge about how to process the information and the site at which such processing takes place. Some account of the nature of memory and of the way in which information is stored, retrieved and manipulated, is clearly essential for any adequate description of the problem-solving process, particularly in a content-rich domain such as chemistry. A full discussion of the role of memory and of the development of an appropriate model for the study will be found in Chapter 2. In this connection, the adoption of a relatively unsophisticated model in the information processing mould facilitated the development of essentially descriptive accounts of candidates' problem-solving behaviour. It was therefore unnecessary, at least in the first instance, to consider the theoretical positions adopted by different schools of psychology towards problem-solving.

(iii) **Cognitive style.** The term "cognitive style" is used to describe a number of relatively stable bi-polar differences between individuals as regards their cognitive functioning. The relevance of research in this field to science and mathematics education has been persuasively argued by Kempa (1979). Such research has, for the most part, focussed on the learning situation and on task performance rather than the processes of solving problems. Although it would have overloaded the present study to introduce this concept, three of the styles identified will be briefly mentioned inasmuch as they throw a sidelight on the sort of individual differences which might be anticipated at certain stages in the problem-solving process.

Work on "field dependence/independence" as a cognitive style has recently been reviewed (Witkin, Moore, Goodenough and Cox, 1977). Field dependent people tend to be global in their perceptions of information and to fix attention on salient cues. Field independent people, on the other hand, tend to be analytical in the way in which they direct their attention. Such individual differences in the perception and apprehension of an examination item might be expected to lead to differences in the initial strategies adopted.

Holzman and Gardner (1972) have discussed a "levelling/sharpening" dimension first described by Gardner et al (1959). This relates to the way in which individuals appear to organise information in long-term memory. Levellers tend to recall rather undifferentiated generalisations concerning memorised information while sharpeners tend rather to recall specific detail. Such differences, which are not related to IQ, might obviously be associated with quite different solution routes on particular examination items.

Finally the "impulsive/reflective" dimension described by Kagan (1966, 1972), appears particularly relevant to problem-solving in the context of examinations (Kempa, 1979). Some pupils tend to be reflective and analytical in their thinking habits, while others of similar ability tend to jump to conclusions. Although relating to modes of thinking rather than to perception, there seems to be some parallel with field dependence/independence. Kagan (1972) has identified two stages in problem-solving at which the impulsive/reflective dimension may act, namely the selection of a method of attack and the evaluation of the solution arrived at. Unfortunately further work on the problem-solving of reflective and impulsive subjects has concentrated on levels of performance on different tasks rather than on details of strategy (eg Zerlinker and Jeffrey, 1975; Rollins and Genser, 1977).

(iv) **Gender.** The negative attitude of many girls towards the physical sciences has been the subject of increasing concern and increasing research in recent years. Ormerod and Duckworth (1975) have suggested that a factor contributing to such attitudes may be differences in cognitive styles between the two sexes. Thus boys may be more able to break "set" (for example in a "hidden figures" test) than girls and tend to perform better on tests of creativity and analytical thinking. Such differences might be reflected in problem-solving behaviour. There might also be differences associated with the tendency of girls towards greater verbal fluency and lower spatial ability than boys. Although it was not deemed appropriate to make this issue a major focus of interest in the present study, it was felt that it would be important not to lose sight of gender as a possible factor influencing problem-solving behaviour and to attempt some comparisons if possible.

1.4 An outline of the approach adopted in the investigation

The first phase of the study consisted of an exploratory empirical investigation, using tape-recorded protocols, into the cognitive behaviour of O-level chemistry candidates on a restricted range of examination items. Follow-up investigations comprising the second phase of the study, and using a different methodology, explored the relationship between candidates' problem-solving behaviour and

(i) the quantity of information provided in an item, and

(ii) the organisation of concepts in a student's long-term memory.

An analysis-of-variance design was employed in the former investigation and cognitive mapping, based on the results of a word association test was used in the latter.

It will be convenient to conclude this chapter with a summary of the way in which the report has been organised. Chapter 2 is devoted to a detailed rationale for the first phase of the study. The nature of problem-solving is considered and the relevant research literature is reviewed. A descriptive model of the problem-solving process is developed and the chapter concludes with a discussion of methodological issues and the presentation of a simple typology of examination items. In Chapter 3 the planning and administration of the first phase of the study are described. The results are then presented and discussed at length in Chapter 4. Candidates' problem-solving strategies and errors are described and a number of broad conclusions are suggested. Chapter 5 presents a detailed rationale for, and describes the conduct of, the follow-up study. After identifying the areas to be addressed, relevant literature relating to the mapping of cognitive structure is reviewed. An analysis of task formulation (in the contect of examination items) is reported and the aims and methods of the second phase of the study are discussed. The chapter concludes with an account of the construction and administration of the instruments employed. Chapter 6 is devoted to the presentation and discussion of results relating to the influence of task formulation (quantity of information) on problem-solving behaviour and Chapter 7 explores results relating to the relationship between candidates' cognitive maps and their problem-solving behaviour. A short summary of the overall findings is presented in Chapter 8, and the implications of these both for classroom and examination practice, and for further research, are discussed.

CHAPTER 2

THE RATIONALE FOR AN EXPLORATORY STUDY

It is usual to precede a formal account of a study of this nature with a fairly extensive literature review. However, although frequent references to relevant work will be made throughout the report, the traditional approach has been modified for two reasons. The first relates to the relative scarcity of publications directly relevant to the first phase of the study and the second to the undesirability of anticipating the results of the first phase (which was essentially exploratory and hypothesis-generating in nature) at this stage of the report. These points are briefly elaborated below.

The first phase of the study was concerned with exploring and describing the cognitive processes involved when students attempt to answer examination items, and this appears to be a relatively new field. The literature of examining offers little that is relevant and, although that of problem-solving is more fruitful, attention has generally been focused on puzzles and games or more recently on mathematical word problems. It is by no means clear how the results of such research can be generalised in content-rich domains such as science. What research there has been relating to problem-solving in science has generally been concerned with the teaching/learning situation and with performance rather than with process. A recent review draws attention to the relatively limited volume of publications on problem-solving in Science Education over a period of 60 years (Champagne and Klopfer, 1977). Articles were classified as concerned with (i) problem-solving ability as an outcome of instruction, (ii) problem-solving as a method of instruction, or (iii) problem-solving as a psychological or sociological phenomenon. Recent work in the information processing mould has started to direct more attention towards the **processes** of problem-solving in science, but Larkin (1980) considers that such work has so far been limited both in extent and impact. It will, however, be necessary to discuss some aspects of the problem-solving literature insofar as they provide part of the context for the present study. This has been done in Sections 2.1 and 2.2 below.

The second phase of the study was planned only after the emergence of certain specific questions as a result of the initial exploratory phase. In preparing this report it was deemed appropriate to adhere to the same chronological sequence. Indeed it would be impracticable to discuss the literature relating to the second phase without reference to the findings of the first. References to such literature will therefore be postponed until the appropriate chapters.

The discussion which follows will attempt to establish the rationale underlying the first phase of the study. In Section 2.1 the nature of problem-solving is considered and the use of this term in the context of the study is explained. A limited review of the literature follows in Section 2.2 and in Section 2.3 a model is developed as a framework against which to describe problem-solving behaviour. Methodological issues are considered in Section 2.4 and the decision to employ protocol analysis is reported. Finally a broad typology of items is discussed in Section 2.5 as a preliminary to describing the conduct of the first phase of the study in the next chapter.

2.1 Towards a definition of problem-solving

In the broadest sense problem-solving describes the behaviour of a subject in achieving or trying to achieve some goal state starting from some different initial state. Before exploring this definition further in the context of the study, it will be useful to consider how the very wide variety of potential problems may conveniently be classified.

2.1.1 The classification of problems. Reitman (1965) has suggested that both goal states and initial states may be well or ill-defined, giving four classes of problems. A slightly different approach to classifying chemistry problems by Ashmore et al (1979) was mentioned in Chapter 1 (Section 1.3.1) and is illustrated in Figure 2.1.

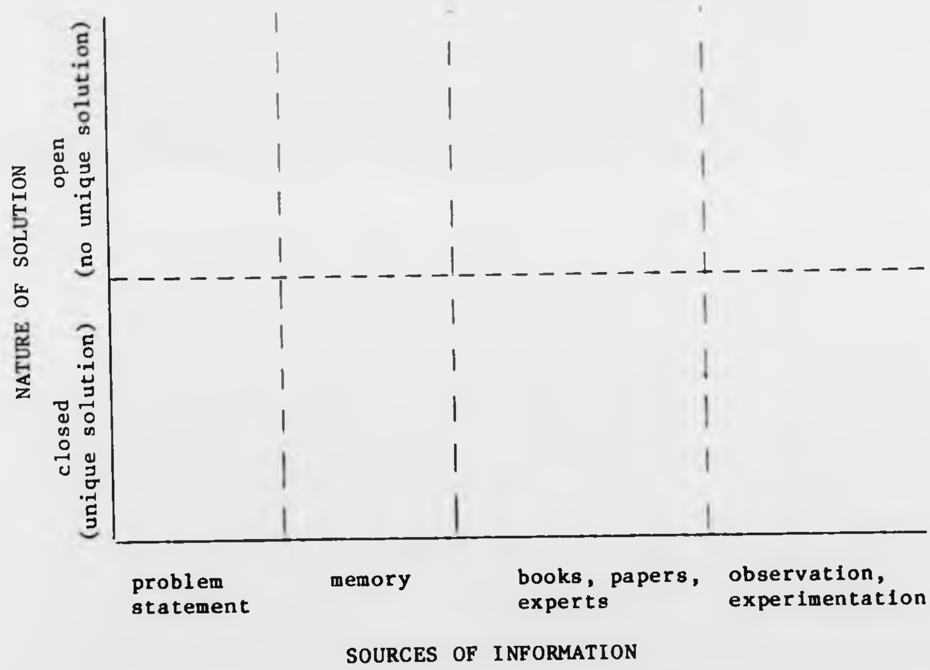


Figure 2.1 A classification of chemical problems (from Ashmore et al, 1979)

The problem solution (goal state) is defined on a closed/open (well-defined/ill-defined) axis and the initial state is represented by a series of possible sources of information. The word "closed" applies to problems with a unique solution such as laboratory analyses and numerical exercises in textbooks, and "open" to situations such as those encountered in fundamental research where a variety of solutions may be acceptable. As regards sources of information these are seen as varying from the strictly limited source represented by the problem-statement itself, through memory and the literature to the potentially unlimited source represented by observation and experimentation. In the context of a written chemistry examination at O-level the goal state is specified in an examiner's marking schedule and is clearly rather well-defined or closed. The initial state is equally well-defined by the information given in the question. The useful feature of the model illustrated is that it directs attention to the role of memory as an additional source of information available to the examination candidate.

2.1.2 **Some perspectives on problem-solving.** Most definitions of problem-solving have insisted that the term be restricted to situations which present an essentially novel challenge to the problem-solver. In trying to summarise such definitions, Mayer (1977) suggests that the problem-solver "does not already know the answer; that is, the correct sequence of behaviour that will solve the problem is not immediately obvious". Although this begs the question of what might or might not be immediately obvious to any given individual facing any particular situation, this type of definition is often countered. For example Bloom and Broder (1950) regard a problem as "a task for which a subject does not have an immediate solution" and Polya (1968) suggests that problem-solving is based on cognitive processing that results in "finding a way out of a difficulty, a way round an obstacle or attaining an aim that was not immediately attainable". In the information processing field a problem has been said to exist when a person "... wants something and does not know immediately what series of actions he can perform to get it" (Simon and Newell, 1972).

A view of problem-solving which places a central role on "insight" in bridging the gap between givens and goals is to be found in much earlier work belonging to the Gestalt school of psychology. A fundamental distinction is made between the "productive" thinking involved in "true" problem-solving and thinking of a merely reproductive kind (eg Maier, 1930; Wertheimer, 1959). Productive thinking is described in terms of a problem-solver's novel reorganisation of structural elements in the problem situation to achieve a solution. Such reorganisations are recognised as "insights" and often occur suddenly following a period of "incubation". The "intuitive" mode of problem-solving described by Bruner (1963) is probably related to the insights of the Gestaltists; however, he has also drawn attention to the existence of analytical modes. He contrasts the "immediate apprehension" of the former with the logical, stepwise, deductive sequence of the latter.

The new learning associated with the solving of problems is central to the differing views of Gagné (1970) and Ausubel (1978). Problem-solving stands at the apex of Gagné's hierarchy of stages of learning and involves the discovery of new combinations of previously learned rules. He deprecates the use of the term in situations which demand only the application of such rules in more familiar situations and suggests that problem-solving performance is related to individual differences in (a) the amount of information stored in memory, (b) ease of recall, (c) concept distinction, (d) fluency of hypothesis generation, (e) retention of the solution model and (f) matching instances to a general class in the verification stage. Ausubel (1978) describes problem-solving as "..... any activity in which both the cognitive representation of prior experience and the components of a current problem situation are reorganised to achieve a desired objective".

Despite this emphasis on "meaningful" learning (ie. on the modification of the existing cognitive structure) Ausubel explicitly allows that the transfer of principles to new variants of familiar problems comes within his description. He notes however that much of what passes for problem-solving is what he calls "rote discovery learning" exemplified by the "type problem approach" to the teaching of mathematics and science. In elaborating upon Ausubel's approach, Novak (1976) emphasises that problem-solving ability is concept specific and that there can be no general problem-solving strategy except insofar as all problem-solving is actually a process of meaningful learning.

Finally, the information processing view uses the digital computer as an analogue for human problem-solving which is regarded as the stepwise processing of information in such a way as to reduce the gap between givens and goals.

Problem-solving research in the field of science education (reviewed in Section 2.2 below) has generally assumed existing definitions of problem-solving, stressing the novel demands made on the problem-solver often in the context of discovery learning. However, an unpublished survey of the views of 12 experienced chemistry teachers led to the following definition:

> "Problem-solving involves a situation, usually in a novel setting, when a given set of data has to be manipulated in order to obtain new information by applying a generalised rule or principle."
>
> (Ashmore 1976, p 76)

This definition appears to reflect a Gagnéan view insofar as it involves the application of previously learned rules. However, the novelty of the situation was not regarded as an essential characteristic, and it may be that teachers perceive problem-solving as a more routine activity than do researchers.

2.1.3 **Problem-solving in the context of the study.** As already mentioned, traditional ideas relating to problem-solving have been developed largely in the context of relatively content-free puzzles such as rearranging patterns of matches, turning coins and crypt-arithmetic.

Indeed such problems offer obvious advantages for cognitive research insofar as knowledge of subject matter is apparently eliminated as a variable. Science examinations, however, represent a content-rich domain and a candidate's attempts will certainly depend strongly on his store of relevant knowledge and his ability to use it, in addition to any more general cognitive skills. In attempting to consider the nature of problem-solving in this context, there are difficulties in maintaining the distinction between productive and reproductive behaviour. At a practical level the importance of knowledge makes it hard to decide what constitutes a problem situation for any particular candidate. What is novel to one, may be familiar to another and perceived as having certain familiar features by a third. In a content-rich situation the distinction between productive and reproductive problem-solving coincides closely with that between Bloom's (1956) descriptions of his familiar "application" and "comprehension" levels. Yet that author admits the difficulty of distinguishing between items operating at these two levels, referring to the importance of the previous experience of the candidate; in practice this is rarely known in sufficient detail to enable such a distinction to be made with any confidence.

There are, however, more fundamental reasons for regarding the distinction between productive and reproductive problem-solving as inappropriate. As a teacher, the author was only too familiar with the situation in which many pupils appeared to struggle with quite simple examination items which ought to have been well within their compass, even items apparently testing recall or the ability to use a simple calculative routine. In similar vein, a number of master's degree studies at the University of East Anglia have suggested that pupils may fail to solve problems of a familiar kind although they appear to possess all the necessary knowledge (Powell, 1976; McCabe, 1977;

Wallace, 1978; Khan, 1980). More recently Massey (1982) has drawn attention to what he calls the exposure-mastery gap. In evaluating 16+ examination results in chemistry he found that many items apparently testing only simple recall or low levels of comprehension were not mastered by candidates passing with good grades. It seems clear that some test items may be presenting greater problems for candidates than teachers or examiners have anticipated. And indeed it is probably misleading to speak of recall as "simple". The reconstructive nature of memory has long been known (Bartlett, 1932) and more recent studies in the information processing paradigm have drawn attention to the use of propositional knowledge in reconstructing memorised information (Norman, 1973). In developing an information processing model of learning and problem-solving in the context of university level chemistry Atkin (1978) suggests that reproductive and productive problem-solving do not differ in kind, but merely in the level of demand which they place on cognitive control processes. Lindsay and Norman (1977) have recorded explicitly that "the person recalling material seems to be solving a problem" and Greeno (1980b) had recently called for the abolition of the distinction between the reproduction of knowledge and problem-solving on the grounds that (i) it is increasingly apparent that knowledge plays a huge role in all problem-solving, and (ii) performance in recall and apparently routine tasks shows the same essential characteristics as "real" problem-solving. This point of view was adopted in the present study and the term problem-solving (PS) will be used to denote the cognitive behaviour of examination candidates in attempting to answer any O-level chemistry item.

2.2 A review of relevant research into problem-solving

The limited impact of research in PS upon science education has been noted in recent reviews (eg Belanger, 1969; Shann, 1976; Ashmore, 1976;

Larkin, 1980; Mujib, 1980). Even in mathematics where word problems have received some attention Hollander (1978) speaks of few studies and limited conclusions regarding the thought processes employed by students. Given this fact and the vast, disorganised nature of the literature of PS as a whole (Shann, 1976) the present review will be highly selective. Before proceeding, therefore, an attempt will be made to locate the study in the wider context of PS at large.

It has already been pointed out that PS in a content-rich domain such as science will differ in obvious ways from that in the content-free tasks typical of much research, in particular as regards its dependence on an organised body of memorised knowledge (Ashmore, 1976; Ashmore et al, 1979). It will also be relevant to consider the age and maturity of the students involved. The PS of O-level candidates will have relatively little in common with that of primary school children, particularly in science. Not only is the science knowledge-base of the latter very limited, but their lower levels of cognitive development and more concrete modes of thinking make it unlikely that findings could be generalised from one age group to the other (Inhelder and Piaget, 1958; see also Shayer and Adey, 1981 concerning the proportions of pupils at various Piagetian stages at different ages). Thus the review will focus on studies relating to science, and where relevant mathematics, at the secondary level and upwards. Finally, consideration must be given to the nature and focus of interest of research within the field of secondary (or higher) science education. The study is concerned with the **processes** of PS in the context of examinations, while much of the literature relates to PS **performance** and to the context of teaching and learning (Champagne and Klopfer, 1977).

In attempting to give structure to a similar review in the context of work on the National Science Foundation's Unified Science and

Mathematics for Elementary Schools Project, Shann (1976) suggested the following categories:

(i) Problem-solving in philosophy

A. Problem-solving as logical thinking

B. Problem-solving as scientific method

(ii) Problem-solving in psychology

A. Models of problem-solving

(a) Models identifying component intellectual processes

(b) Models focussing on the operation of problem-solving processes (information processing models)

B. Empirical studies

(iii) Problem-solving in science education.

However, it was felt that the various schools of thought in the field could be better reflected by adopting the categories employed in Klein and Weitzenfeld's (1976) review. They refer to:

(i) The Deweyan tradition

(ii) The Gestalt model

(iii) The information processing model.

For the purposes of the present review an additional category of weakly affiliated studies will also be considered.

The Deweyan tradition is firmly training-orientated and, although much concerned with describing stages in PS and offering heuristic advice, lacks any strong investigative or theoretical basis. The Gestalt model emphasises structural aspects of problems and seeks to identity PS processes. Many studies employ essentially clinical research techniques. The third category of literature reviewed is fairly broad. It accommodates a number of studies based on investigative research which have weak theoretical affiliations or which belong to the schools of Gagné, Bruner or Ausubel, all of whom have been influential in the field of science education (Shulman and Tamir, 1972). Investigative paradigms range from the clinical to the truly experimental. Finally, the information processing model focuses on the internal processes of PS using analogies from the world of computing. With its explicit emphasis on the role of memory it appears particularly well adpated to the study of PS in content-rich domains. Although experimental studies are also reported the clinical paradigm has been widely adopted.

2.2.1 **The Deweyan tradition.** The Deweyan tradition, as described by Klein and Weitzenfeld (1976), is firmly training orientated. It depends on the logical analysis of what Dewey would have called "reflective" thinking but lacks any systematic investigative or theoretical basis. Dewey (1910) identified "five logically distinct steps" in PS which have been paraphrased in modern terms by the above authors as:

(i) recognition that there is a problem;

(ii) defining the problem;

(iii) generating hypotheses as to the solution of the problem;

(iv) elaborating these hypotheses and inferring their properties;

(v) testing the hypotheses.

This tradition has been concerned to help to improve PS performance by focussing on a logical sequence of activities and developing accompanying heuristics. A number of recent lists in differeng contexts closely reflect Dewey's original proposals (eg Gagné, 1966; Greeno, 1973; Joseph, 1974; Jackson, 1975; Ausubel, 1978). The popular work of de Bono (1967, 1969) represents a more radical approach in the same tradition.

In the field of mathematics Polya's (1957) popular "heuristic" presents a simplified four-step approach for teachers (and students) of mathematics:

(i) understanding the problem;

(ii) devising a plan;

(iii) carrying out the plan;

(iv) looking back.

In discussing each stage detailed hints of a heuristic kind are provided.

Work reported by Chorneyko et al (1979) in the field of chemical engineering belongs firmly to the training tradition but relates to PS in an industrial rather than an educational context. The rather complex and idiosyncratic approach adopted (not least in attributing the paper to no less than 12 authors!) makes it difficult to transfer the ideas discussed to the school situation. However a set of PS stages developed in the context of university level chemistry (Ashmore et al, 1979) seems likely to be more useful:

(i) define the problem;

(ii) select appropriate information;

(iii) combine separate pieces of information;

(iv) evaluate the solution.

More recently Selvaratnam and Frazer (1982) have published a guide to the solving of numerical problems in general and physical chemistry at upper secondary and early undergraduate levels. The five stage procedure summarised below is accompanied by detailed advice of a heuristic kind in addition to basic factual information and practice exercises. The stages are:

(i) clarify and define the problem;

(ii) select key equations;

(iii) derive the equation for the calculation;

(iv) collect the data, check the units and calculate;

(v) review, check and learn from the solution.

Clearly this example relates to a rather particular range of problems and also reflects the views of at least one of the authors regarding the importance to students of memorising basic equations rather than relying on the recall of propositional information (Selvaratnam, 1983).

In reviewing the influence of Dewey's work on school science education, Champagne and Klopfer (1977) report no attempt to validate his PS stages, however the idea that problem-solving can be usefully described in terms of logical steps has stood the test of time and clearly retains considerable value as a basis for analysing the process.

2.2.2 **The Gestalt model.** The Gestalt model has provided the basis for much of the classical work on problem-solving. It differs from the Deweyan tradition in having both an experimental and a theoretical basis and in emphasising intuitive rather than analytical thinking.

The problems studied by Gestalt psychologists have tended to be content-free rather than dependent on detailed knowledge of a particular subject and are often quite complex. Examples include the construction of a pendulum given a rather unpromising set of materials (Maier, 1930), two-dimensional construction puzzles (Durkin, 1937) and "practical" problems such as planning the irradiation of a tumour without damage to surrounding tissue (Dunker, 1945). The methods of study were essentially clinical and based on detailed observation of individuals as they tackled the problems. Durkin and Dunker also made use of recorded protocols.

The theoretical position adopted in interpreting this and much similar work reflects the view of the Gestalt school on perception. Much attention has been directed upon the subject's understanding of the problem and its demands. Successful PS often involved a period of "incubation" followed by a sudden "insight" in which the structure of the problem was mentally reorganised. Ideas such as "set" and "functional fixedness" were developed to explain the difficulties experienced by problem-solvers in achieving such structural reorganisation, and "direction" referred to the influence of cues which helped to facilitate insights. In general the Gestalt school has shown little interest in problems which can be solved by a routine or analytical approach and indeed contrasts "productive" thinking with lower levels of cognitive behaviour (Wertheimer, 1959).

The theoretical basis of work in the classical Gestalt tradition has been criticised as too vague to be rigorously tested by experiment (Mayer, 1977) and there seems to have been no work relating directly to science education (although Wertheimer (1959) has used the learning of some aspects of school geometry to contrast "structural understanding" with rote memory). Nevertheless, many of the methods employed and the ideas generated continue to have a significant influence on PS research.

2.2.3 **Weakly affiliated studies of problem-solving in science (and mathematics) education.** The studies reviewed in this section tend to be eclectic or to have rather weak theoretical affiliations although a few make particular reference to various well-known schools of thought.

A series of unpublished dissertations relating to work undertaken at the University of East Anglia in the context of master's degree studies are concerned with PS in chemical education at various levels. Although clear findings are limited, some general trends are worth reporting. A number of the studies have employed "PS networks" (Bapat, 1975; McCabe, 1977; Morris, 1977) and these have also been referred to in the published literature (Ashmore, Frazer and Casey, 1979). The networks simply show how information given in the problem and retrieved from memory can be combined to arrive at the solution. Advantages claimed for the network approach refer to various aspects of the analysis of a problem and to the presentation of the information in a single diagram. However, there is little evidence relating networks to the actual solution strategies used by students. Three of the studies were concerned with investigating such strategies, and in particular with identifying students' difficulties, using the analysis of written PS attempts supplemented by questionnaires and/or other test instruments (Powell, 1976; McCabe, 1977; Morris, 1977). Although little was learnt regarding strategies, a number of interesting findings emerged regarding PS failures. All three reported that students' PS attempts frequently broke down through failure to use knowledge which other tests indicated clearly that they possessed. This was brought out particularly well in McCabe's study with GCE O-level students. On average, 41% of relevant knowledge known to be stored in memory was not used, nor indeed was 22% of relevant knowledge given in the questions, leading in both cases to PS failure. Studies by Wallace (1977) and Khan (1980) using clinical methods with A-level students recorded similar findings.

Interestingly, Powell (1976) found that "fluency of chemistry knowledge" as measured by a 100-item multiple choice test, correlated better with PS performance than a standardised measure of reasoning ability (AH5). Other sources of difficulty noted by McCabe and by Morris at O-level, and by Ashmore (1976) with undergraduates included:

(i) misreading or misunderstanding the problem;

(ii) errors in reasoning or in combining information (Morris records that few O-level pupils could successfully handle as many as four processes);

(iii) errors in computation;

(iv) mathematical difficulties relating particularly to proportion and to the mole concept;

(v) failure to check the answer.

In a very recent paper Selvaratnam (1983) classifies A-level and undergraduate students' difficulties as:

(i) content difficulties - that is, difficulties associated with both the selection and the accurate recall of knowledge from long-term memory;

(ii) process difficulties - that is, difficulties in clarifying the problem and defining the goal, and in deciding where to start; and failure to proceed systematically and to use equations.

It is of interest to note that Hollander (1978), in summarising her review of research on the thought processes involved in the solution of verbal arithmetic problems, criticises the tendency of authors to list errors rather than to concentrate on strategies and on the causes of errors. The same could be said of much of the work referred to above.

The direct investigation of the PS process through protocol analysis has generally been associated with either the Gestalt school or with work based on the information processing model. However, there have recently been a few studies in science and mathematics education without these affiliations. Three of the above PS studies from East Anglia made some use of the technique (Bapat, 1975; Wallace, 1977; Khan, 1980), and although two added little to what has already been recorded, Bapat's study was certainly more significant. He explored the responses of graduate and post-graduate students and of university staff to difficult problems in chemistry, and supplemented the protocol analysis with a test of prerequisite concepts. The protocols were converted into "strategy trees", 95% of which were found to be consistent with Gagné's learning hierarchies (Gagné, 1970). He also reported favourably on the use of protocol analysis for work of this sort and noted that 90% of the subjects used an analytical as opposed to an intuitive approach. It may be, of course, that these results were influenced by the unusually high academic level at which the work was conducted.

Three unaffiliated studies are principally of interest in regard to their use of a checklist or tally sheet to simplify data obtained by protocol analysis. In a pilot study in which post-graduate students answered fixed response chemistry items, Hateley (1979) categorised responses as belonging to one of eight pre-determined classes. These related both to the correctness of the response and to the reasons given for selecting it. The latter included, for example, "reason given is same as examiners", "other legitimate reason", "memory" and "guess". The method appeared to provide useful if limited information regarding, for example, the proportion of the students obtaining correct responses by the route intended (68% in this study), by memorisation alone (4%), by guessing (3%), or by incorrect reasoning (4%). The pilot study has not been followed up. More elaborate checklists have recently been used in analysing PS protocols in the field of mathematics (eg Mandell, 1980; Webb, 1979). Mandell developed a "Problem-Solving Behaviour Tally Sheet" on an eclectic basis drawing on "several theoretical studies in the literature". The responses of sixth-grade pupils on a mixed set of mathematical word problems and general puzzles were classified as belonging to one of two to four categories during each of four phases of the PS process. This facilitated a numerical description of the results, however the categories employed are of very limited relevance to the present study. Webb (1979) coded the protocols of high school pupils on simple algebra problems using a highly elaborate checklist of PS variables, although these are by no means adequately explained in his paper. Of more interest were the way in which this facilitated the numerical handling of the results (factorial and regression analyses were employed), and the conclusion that 50% of the variance of the pupils' PS scores could be attributed to "conceptual knowledge" and only 13% to "process components".

Protocol analysis has also been used to explore PS behaviour in a more open-ended and descriptive way. Bloom and Broder (1950) studied students' attempts at university examination questions. Their analysis of the data obtained suggested that four critical areas were associated with PS performance:

(i) understanding the nature of the problem;

(ii) understanding the ideas contained in the problem;

(iii) general approach to solution of problems;

(iv) attitude towards the solution of problems.

The extensive work of Krutetskii (1976) in mathematics, and particularly with mathematically gifted children, is well known. On the basis of his analysis he identified the following structure of mathematical abilities in school-age pupils:

(i) obtaining information - grasping the structure of a problem;

(ii) processing information -

(a) logical thinking,

(b) generalisation,

(c) curtailment (ability to take short cuts),

(d) flexibility,

(e) elegance (clarity, simplicity, economy, rationality),

(f) reversibility;

(iii) retaining information - memory for mathematical relationships, characteristics, arguments and proofs, PS methods and principles of approach;

(iv) general synthetic component - mathematical cast of mind.

Reif, Larkin and Brackett (1976) found tape-recorded protocols useful for analysing the PS processes of college physics students prior to devising materials to teach PS skills. They concluded that, before such instruction, problems were often tackled "in very haphazard and ineffective ways". And in the field of mechanical engineering, Cowan (1977) used students' "running commentaries" to compare and classify their PS methods. The results were very specific to the type of problems investigated but the author was convinced of the value of protocol analysis for this sort of study and drew attention particularly to the possibilities it might open up for matching instructional approaches to the PS styles of individual students.

2.2.4 **The information processing model.** The information processing approach attempts to describe human PS in terms of computer processing. A simple information processing model is illustrated in Figure 2.2.

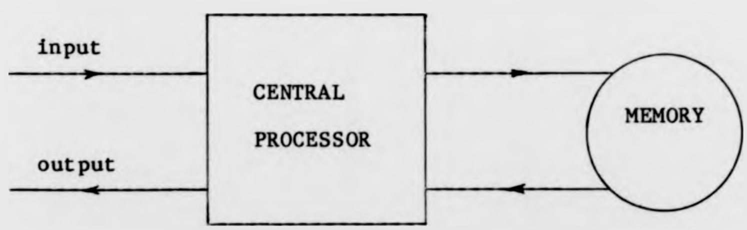


Figure 2.2 A simple information processing system (adapted from Simon and Newell, 1972)

"Input" to the central processor corresponds to the subject's perception of the problem and "output" to his response. The central processor also draws on information from memory and operates selectively on inputs from both sources to produce a sequence of differing information states until a solution is reached or the problem is abandoned. The selection of operations which tend to reduce the gap between givens and goals is controlled by "means ends" analysis. A fairly detailed account at an elementary level is given in Lindsay and Norman (1977) or a fuller one in Simon and Newell (1972). Work within this paradigm can be broadly divided between that which is explicitly linked to ideas of computer simulation and of artificial intelligence, and that which rather regards the functioning of the computer as providing a useful analogy against which to analyse human cognitive behaviour. The information processing approach has been criticised for its failure to consider the identification and definition of problems (Klein and Weitzenfeld, 1976) but its strength lies in the discipline it brings to speculation regarding the **processes** of PS. However the human/machine analogy is no more than an analogy. Computer simulations mirror human behaviour only imperfectly and cannot, for example, engage in "productive" thinking. And as Mayer (1977) has pointed out, even where overt human behaviour is well simulated this does not mean that the underlying data processing is the same.

The attraction of the information processing model as regards work on PS in the context of school science lies not only in its sharp focus on process, but also in its emphasis on the role of memory which obviously plays a vital role in such a content-rich domain. A series of articles by Greeno (1973, 1976, 1978, 1980a, 1980b) have examined the role of memory in mathematics education and a recent paper by Atkin (1978) describes a memory model developed to account for learning and PS in the context of chemistry. Both will be referred to later when the development of a PS model for the present study is discussed. Meanwhile, relevant experimental and clinical work in both mathematics and science is reported below.

Experimental studies relating to the information processing model seem to have payed more attention to mathematics than to science. Of particular interest is a series of studies with university mathematics students tending to link learning, cognitive structure and PS behaviour (Mayer and Greeno, 1972, 1974; Mayer, 1975). Treatment - post-test interaction (TPI) and aptitude - treatment interaction (ATI) studies supported the idea that PS behaviour, especially as regards near and far transfer, can be explained in terms of the internal organisation of long-term memory which in turn reflects different levels of teaching and learning. Reception learning was associated with the addition of unrelated items of information to memory and with reproductive behaviour (no transfer). Teaching which emphasised algorithms and formulae, and linked these to prerequisite knowledge in the manner proposed by Gagné (1970), was associated with high levels of "internal connectedness" in cognitive structure (that is, with a well integrated body of information which might, however, be isolated from previously existing information structures) and with good near transfer. Finally, teaching for "meaning", emphasising links to past experience in the manner proposed by Ausubel et al (1978), was associated with high levels of "external connectedness" (that is, with information which was well integrated into the existing cognitive structure) and with good far transfer. In addition to linking explicit features of memory organisation to PS behaviour, the eclectic element relating such features to the work of Gagné and Ausubel is of interest, calling to mind Shulman and Tamir's (1972) plea for a synthesis of ideas as regards psychological approaches to science education. Indeed Mayer claims elsewhere that the information processing paradigm

> "..... offers a breakthrough in the psychology of thinking which may ultimately produce a precise reformulation and interpretation of Gestaltist, associationist, and other ideas."
>
> (Mayer 1977, p 143)

An experimental study relating to science has been reported by Atkin (1978). This was concerned with the validation of the author's model of learning and PS, which combines the information processing approach to memory and problem-solving with the ideas of Ausubel. The model has many features in common with that underlying the Greeno and

Mayer studies and will be referred to in more detail in Section 2.3 below. Meanwhile it can be recorded that a TPI design using university chemistry students yielded results consistent with the model.

In the field of mathematics an experimental study by Malin (1979) draws attention to a different issue. He examined the efficiency of predetermined PS strategies on a set of algebra problems with junior high school pupils. Most of the findings were of rather limited general interest. However it was clear that situations involving high processing loads, that is, which demanded the simultaneous handling of several items of information, were associated with a decline in PS performance.

A review of four studies based on a clinical paradigm and employing protocol analysis will complete this section. Bhaskar and Simon (1977) report a study of the PS behaviour of a single subject in the field of chemical engineering thermodynamics, using detailed analysis of "thinking-aloud" protocols. They draw attention to the subject's use, in a semantically rich domain, of memorised information in the form of both "data structures" and procedures, as well as general PS approaches, particularly means-ends analysis, which are not domain-specific. PS errors were found to lead to deviations from the procedures generally followed by the subject and to the injection of new strategies.

The PS behaviour of an expert and that of a novice have been compared in two studies in the field of college physics (Simon and Simon, 1978; Larkin and Reif, 1979). In a useful introduction, the latter authors draw attention to the need for a relevant information processing model of PS. They suggest that this might be constructed either through logical analysis or by observation of PS in progress. They further suggest that PS mechanisms are unlikely to be revealed by statistical studies and advocate concentration on individual cases using protocol analysis, PS steps being identified by an iterative process. Both the studies referred to reported similar findings. The experts tended to seek overall understanding of the problem and to plan a coherent route to the solution, while the novices tended to write a series of equations to link the givens and goals without prior planning.

The last study reported here also relates to physics, this time at undergraduate level (Mujib, 1980). The author focused most of his attention on the difficulties of analysing the protocols which he obtained by interviewing students concerning PS episodes. He developed a method based on network analysis (Ogborn, 1979; Bliss and Ogborn, 1979) which will be discussed later in this report.

2.3 A descriptive model of the problem-solving process

A theoretical model is obviously helpful if PS behaviour is to be described in in a systematic way. The model must provide a framework for describing the **sequence** of mental activities associated with the process and must also reflect the important role played by memory in content-rich domains. Since the aim of the first phase of the study was simply to observe and describe PS behaviour, it was also important that the model should, as far as possible, be neutral as regards the adoption of any particular theoretical stance. A composite model described by Greeno (1973) was selected as meeting the necessary conditions and was modified slightly on the basis of (a) the literature reviewed above, (b) iterative inspection of a range of typical GCE O-level chemistry examination items and (c) preliminary observations of candidates' PS behaviour (see Chapter 3). It consists of a listing of PS stages firmly based in the Deweyan tradition (see in particular the stages identified by Polya, 1957; Gagné, 1970; Ausubel, 1978 and Ashmore et al, 1979) associated with a simple memory model and is illustrated below in Figures 2.3.1 and 2.3.2.

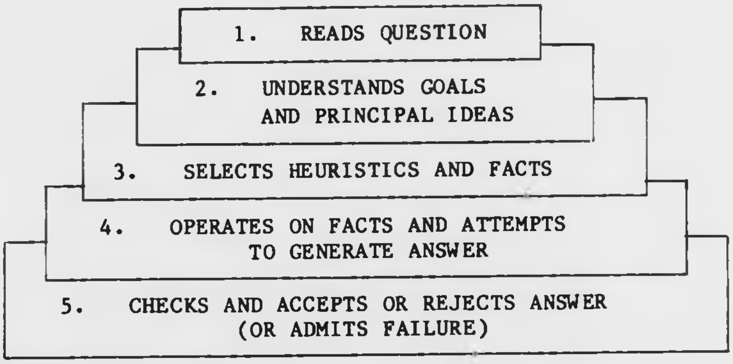


Figure 2.3.1 Steps in anwering examination items
(adpated from Greeno, 1973)

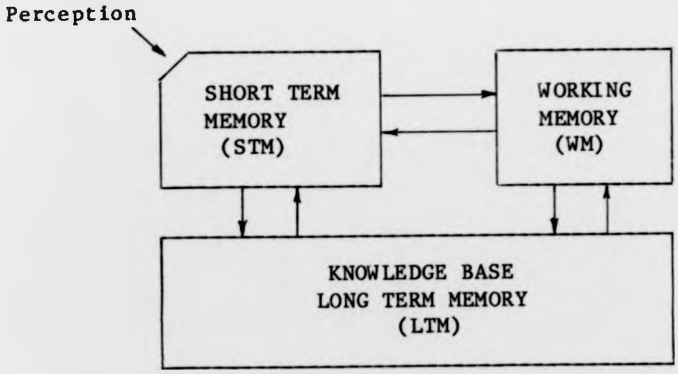


Figure 2.3.2 Structural components of memory
(adapted from Greeno, 1973)

It will be convenient to discuss the memory model before elaborating on the PS steps.

2.3.1 A simple model of the structural components of memory. Larkin and Reif (1979) caution against the temptation to over-elaborate a model of this kind. The aim is simply to provide a framework against which to interpret and describe empirical observations of students' PS behaviour. The study was to be broad and exploratory in nature. It was in no way an attempt to validate any particular theory but rather to generate hypotheses concerning pupils' success or failure in solving examination items. The model suggests that information from such an item enters a student's short-term memory (STM) and is passed on to working memory (WM). Here, together with information from his knowledge base in long-term memory (LTM) store, it is processed in an attempt to generate an answer. The characteristics of each of these three components are discussed below.

(i) **Long-term memory** Long-term memory (sometimes referred to as semantic memory) represents an individual's accumulated knowledge-base. It can be regarded as mediating between the students' learning experiences and the outcome of such learning (Mayer and Greeno, 1972; Mayer, 1975; Gagné and White, 1978). Although the details are not important in the present context, it is assumed that the information in LTM is highly structured and can be represented by a network of labelled associations (Rumelhart, Lindsay and Norman, 1972; Simon and Newell, 1972). Within this structure Lindsay and Norman (1977) have found it useful to distinguish three types of information employed in PS, propositional, algorithmic and heuristic.[1] Propositional knowledge refers to knowledge of facts, concepts, principles and theories, while algorithmic and heuristic knowledge is concerned with ways of processing information.[2] Algorithms are sets of rules or procedures which, properly applied, will automatically generate a correct answer. Heuristics on the other hand, are simply rules of thumb or general plans of action which are often useful in tackling particular types of problems. Although the capacity of LTM is vast and all nodes in the information

---

1. Gagné and White (1978) have proposed a model of LTM in which can be distinguished (a) networks of propositions, (b) intellectual skills (algorithmic and heuristic), (c) images and (d) episodes; however the two latter appear to be of only limited importance in the context of examinations.

2. Both algorithmic and heuristic knowledge can, of course, be expressed in propositional terms; the distinction is essentially one of convenience.

network are in theory addressable, it is not necessarily easy to retrieve any particular item of information. Ease of recall is related both to recency of learning and to the frequency with which the information is used. It also depends upon both the way the information is organised in LTM and the nature of the cues employed to recall it. In the PS situation there may also be difficulties in deciding what information is appropriate to the task.

(ii) **Short-term memory** Short-term memory transforms, interprets and encodes incoming data. It may ignore some information and is selective in what is attended to. It interacts (directly or through working memory) with LTM, storing and retrieving data. STM has a strictly limited capacity as regards both the time for which information can be retained and the amount of data which it can handle. The former limitation appears to be the less serious in that it can be, and habitually is, extended by rehearsal (internalised repetition). A more serious difficulty as regards PS is the inability of STM to handle more than about 7 (±2) items of information at a time (Miller, 1956). The practical limitation, of course, is defined by how much information is represented by each item or "chunk". It is well established that the average person's STM capacity may equally well be 7 randomly selected letters of the alphabet or, say, 7 familiar four-letter words comprising 28 letters. Clearly by encoding and labelling each four-letter group as a single chunk (word) the total amount of information that STM can handle has been significantly enhanced. However Simon (1974) has suggested that STM span gradually falls as the size or complexity of the chunk is increased. He offers a useful, if tautologous, definition of a chunk as "that item of information that the STM can hold about five of"!

The relevance of STM span to PS behaviour is clearly exemplified in two studies already referred to where a high processing load was associated with a falling-off in PS performance (Morris, 1977; Malin, 1979). Individual variations in STM capacity (at most 30% to 40%) cannot alone account for the very wide range of PS performance observed (Gregg and Farnham-Diggory, 1975). However, a series of studies at the University of Glasgow suggests that ability to integrate information into meaningful chunks may have a crucial influence on performance in chemistry (Johnstone and Kellett, 1980; Johnstone, 1981; Johnstone and Letton, 1982).

(iii) **Working memory** Since the early 1970s the two-component view of memory has come to be regarded as an over-simplification (Baddeley, 1976). Working memory, though closely associated with STM, has a considerably greater capacity for holding information and can hold it for longer periods. However its capacity for information processing under conscious control is limited by STM span. It is within WM that data from STM and LTM interact during PS and that a solution route is generated. It is a feature of Greeno's model that the information retrieved from LTM will include sets of "transformational relations" appropriate to the problem as encoded in WM; these may be propositional or algorithmic (or, initially, heuristic).

2.3.2 **Steps in answering examination items.** The PS stages shown in in Figure 2.3.1 are listed in a more detailed way below. It is envisaged that the student:

1. reads the question;
2. relates it to his existing cognitive structure ie understands -
   - 2.1 the nature of the goal,
   - 2.2 the principal ideas in the question;
3. selects knowledge to start from in the form of -
   - 3.1 a "fact" or a "concept",
   - 3.2 a propositon, algorithm or heuristic;

4. operates on selected knowledge (recalling/selecting additional information from the item/LTM as needed/available);

5. checks result by -

   5.1 checking whether results meet goals,

   5.2 checking for recognition (if familiar) or for congruence with existing relevant knowledge,

   5.3 checking operations for accuracy (or using alternative operations).

The main stages 1 - 5 are to be regarded as sequential. However the overlapping boxes in Figure 2.3.1 are intended to reflect the idea that an earlier stage may often be returned to. Thus although stage 4, for example, will be reached only after passing through the three earlier stages, the processing of facts may often lead to a new understanding of what the question involves and/or to the selection of a new method of attack. It is not, however, suggested that the sub-stages (2.1, 2.2, etc) shown in the listing above are necessarily sequential.

The various steps will now be discussed more fully.

(i) **Stages 1 and 2 - "reads" and "understands" the item** Although the reading of the question is a logically necessary first step, the students' understanding of the problem is of more interest. Understanding the ideas in the question and understanding the goal are two of the five parameters used by Bloom and Broder (1950) to distinguish between successful and unsuccessful problem-solvers. Unless the principal ideas in the question are at least in some degree meaningful to the student, progress is unlikely. Of equal importance is the clear identification of the goal. Obviously the way in which these two elements are grasped and related is critical to any attempt to generate an answer. Krutetskii (1976), describing the processes of mathematically capable pupils in grasping problems, speaks of the analysis, reorganisation, translation and encoding of problem elements and Larkin (1980) has drawn attention to the differing "problem representations" of novices and experts in solving physics problems. Expert solvers translate the problem into a qualitative representation while novices start processing without any overall representation. Another characteristic of good problem-solvers is reported to be their ability to relate a new problem to previously solved problems of similar structure (Silver, 1979). Failure to clarify (understand) the problem has been blamed for poor performance on numerical questions in A-level chemistry (Selvaratnam, 1983). This may be linked to Johnstone's (1981) interesting hypothesis regarding students' perceptions of problems in chemistry. High levels of information content in the problem and low-levels of concept development in the student are linked to high levels of perceived difficulty. Students whose relevant conceptual framework in LTM is more highly developed are better able to chunk the information provided and thus avoid overloading the STM/WM system. In his original presentation of the model under discussion, Greeno (1973) **assumed** that the problem-solver constructs a network relating elements in the problem; one suspects that some students may not get that far!

(ii) Stage 3 - Select a starting point and a method of attack

This stage of the PS process is obviously closely linked to the previous stage and to the stage which follows it. Selection of a starting point and a method of attack may be almost synonymous with understanding the question, and it seems likely that the initial stages of operating on the facts will sometimes be rehearsed as part of the selection process. Except in rather straightforward cases it is anticipated that the stages "select" and "operate" (or possibly "understand", "select" and "operate") will be cyclic. Thus it is envisaged that the result of any operation will be evaluated in terms of progress towards the desired goal (means-ends analysis) and that this in turn will lead to further processing using the intermediate result as a new starting point or indeed to a completely fresh start. Clearly, different starting points may be possible including starting from the goal and working backwards and starting from associated information obtained from LTM. Not knowing where to start and failure to select appropriate knowledge and methods from LTM are amongst the main PS difficulties referred to in Selvaratnam's (1983) paper and it is of interest that Bloom (1956) makes the ability to select appropriate principles in a novel situation the key to distinguishing between answering test items at the level of "application" rather than merely "comprehension". Information processing models of LTM such as that employed in this study, envisage some degree of association between conceptual (factual) and procedural knowledge.[1] Thus failure to select appropriate procedures may well be related to failure to locate and/or select appropriate knowledge even when this is known to be present in LTM (as in the findings of Bapat, 1975; McCabe, 1977 and others in the UEA school). This in turn may be due to failure at the "understanding" stage to correctly encode the problem situation. In such cases the approach selected may involve the application of general heuristics (such as reorganising the information given or drawing a diagram or attempting to represent the information given symbolically in the form of an equation) in an effort to reach a better understanding of the problem; alternatively the problem-solver may proceed to generate and test hypotheses or to employ guesswork and/or trial and error. Although obtaining information regarding a student's understanding of a problem seemed likely to present difficulties, it was anticipated that

---

1. The nature and extent of such associations will presumably reflect the students' learning experiences and will certainly influence the likelihood of a particular method being selected for any given PS task. Mayer's (1975) paper relating the "internal/external connectedness" of procedural knowledge in LTM to near/far transfer is particularly relevant.

information about the selection of starting points and procedures might fairly readily be obtained either directly, or by inference, from tape-recorded protocols.

(iii) **Stage 4 - operates on knowledge** This stage refers to the information processing which follows the selection of a starting point and a method of attack. The knowledge selected as a starting point may be transformed by combination with other propositional knowledge or by the application of an algorithm in an effort to reach the goal state. If an approach of this sort has not been selected then information may be transformed by the application of general heuristics including guessing or trial and error in an attempt to make progress; alternatively more insightful hypotheses may be generated and tested. Speculation regarding errors of processing, that is mistransformations of information, suggest the possibility of (a) the use of defective propositions or algorithms, or (b) mistakes in the application of the method selected. The latter might be due to carelessness (that is failure to devote adequate attention to the process) or to overloading of STM. In the context of chemical PS, the errors in reasoning and in calculating referred to by both McCabe (1977) and Ashmore (1978) are presumably processing errors, although sufficient information is not provided to enable the cause to be described more precisely. The sort of errors associated with STM overload have already been discussed (eg Morris, 1977; Johnstone, 1981).

(iv) **Stage 5 - checking the result** The elaboration of the stages suggests that three forms of checking might take place. Although the planned empirical investigation will look for evidence of all or any of these, only the first comes close to being a logical necessity.

Step 5.1 refers to checking the result of processing against the goal state. If a satisfctory match is achieved the result is entered as the answer, otherwise the problem-solver recycles to stage 3 ("select") or even to stage 1 or stage 2. Failures at this point would be associated with inadequate characterisation of the goal or the result.

Step 5.2 refers to checking that the result appears sensible. The generation of the answer may cue its recognition as something the problem-solver already knows or has seen before. Alternatively it may be possible to judge the answer as likely or otherwise by evaluating it in terms of existing knowledge of the appropriate domain in LTM. The well known tendency of students to present obviously ridiculous answers, especially to numerical problems, suggests that this form of checking is far from universal!

Step 5.3 refers to checking of a more mechanical kind where an answer is confirmed by working through all or part of the problem a second time to make sure that errors have not been made. This form of checking might follow exactly the same route as the original solution or might employ an alternative method. Teachers and examiners are only too well aware of the number of mistakes which could be easily corrected if students employed this form of checking. When McCabe (1977) refers to "failure to check answer" (22% of the "mechanical" errors she observed in the PS of O-level chemistry students were so classified) it is presumably this sort of checking to which she refers.

2.3.3 **Atkin's model of problem-solving.** In a paper which became available only after the first phase of this study had been completed, Atkin (1978) has developed a model with many similar features to the one presented here. Her successful experimental validation of the model with university chemistry students has already been referred to. An interesting feature is its explicit fusion of Newell and Simon's (1972) ideas of PS, Lindsay and Norman's (1977) model of human memory and Ausubel's (1978) theory or meaningful learning. It also suggests that the processes of learning and problem-solving have much in common. It differs from the model used in the present study principally in its rather more detailed attention to the minutiae of cognitive functioning, and in its appeal to the roles of "monitor" and "interpreter" rather than to the roughly equivalent regions of cognitive activity labelled STM and WM. The PS sequence described by Atkin is as follows:

1. problem information is input to system;
2. monitor cues system to everyday usage resulting in semantic comprehension;
3. monitor cues system to relevant area of LTM from which information containing related concepts is retrieved and a "problem-space" is set up;
4. interpreter identifies appropriate rules of inference;
5. information from the task and LTM is manipulated to arrive at a solution.

Although providing a more detailed account of the "understand" and "select" stages, Atkin's model is entirely compatible with that already developed. It adds little in the context of the initial exploratory phase of the investigation.

2.4 Methodological issues

The selection of a method of attack for the first phase of the study did not present any great difficulty. Four possible approaches are briefly evaluated below: the experimental validation of a PS model, the analysis of students' answer scripts, the use of interviews and questionnaires and the analysis of tape-recorded protocols.

2.4.1 Experimental validation of PS models.

The use of experimental studies in validating PS models has been supported by Gagné and White

(1978) and examples already referred to in the field of mathematics and science education include work reported by Mayer (1975) and by Atkin (1978). However, this form of validation often appears weak inasmuch as results which are found to be consistent with a particular model are not necessarily inconsistent with equally plausible alternative explanations. In any case the use of such an experimental design can at best provide evidence for or against existing descriptions of PS behaviour while the aim of the present study was the generation rather than the evaluation of such a description. As already explained, the model presented earlier in the chapter was intended merely to provide a general framework, consistent with existing knowledge about PS, against which to describe and interpret observations of PS behaviour. It was not by any means either sufficiently definite or sufficiently detailed to be a subject for validation.

2.4.2 **Analysis of answer scripts.** The analysis of examination scripts or any other written answers to problems assumes that the mental processes of students can be inferred from what they write. This assumption would be hard to justify and Bloom and Broder (1950) have expressed the opinion that working from the products of thought in this way leads to research which is "inappropriate, superficial and, in all likelihood, incorrect". The use of specially constructed problems (Krutetskii, 1976) and the analysis of answers in the light of previously constructed networks (McCabe, 1977) have been attempted as a way of overcoming this difficulty. Such constructions, however, may distort the problem situation, and the theories underlying them strongly influence the way in which results are interpreted. Krutetskii's most interesting findings stem from his use of protocol analysis and McCabe was able to do little more than classify types of errors. The dangers of making inferences from written answers alone, even where students have been asked to show their "working", is well illustrated in a report by Cowan (1977). In attempting to evaluate and improve the performance of a structured learning package in mechanics, he asked students to supplement their written attempts at certain items with a tape-recorded running commentary of their efforts. He writes:

> "But from the ..... recordings the writer was about to receive the greatest surprise of his teaching career. Solutions which seemed on a visual examination to be the miserable products of ill-prepared and stupid minds emerged as commendable records of tortured battle-grounds on which the student, having made one not unreasonable error in his otherwise logical chain of solution, had struggled on manfully, and even logically, until he became more and more bogged down through the consequences of that solitary error. And as if this revelation was not enough, the writer was also astounded to discover how many of the students, good and bad alike, chose not to follow the carefully structured and firmly recommended method of solution which had been the basis of formal instruction."
>
> (Cowan 1977, p 5)

2.4.3 **Use of interviews and questionnaires.** Interviews and questonnaires have generally been used only to obtain supplementary information about PS processes. Although Mujib (1980) used interviews following PS episodes as his principal source of data, it seems improbable that students can be relied upon to recall the processes they used accurately or in detail, and there is a danger that their solutions may be rationalised and idealised. This is particularly the case where students are required to attempt a series of items in an examination context, rather than the single problems employed by Mujib. There is also the problem that good interview schedules or questionnaires are notoriously difficult to design and tend to reflect unduly the preconceptions of their authors. It is hard to envisage an instrument which would be likely to provide adequate information.

2.4.4 **Analysis of tape-recorded protocols.** Inspection of the literature already reviewed in this chapter indicates that protocol analysis must now be regarded as the method of choice for studies of the processes involved in PS. A subject is asked to "think aloud" as he attempts to solve the problem and his verbalisations are recorded for subsequent analysis. The method has been in use for a long time and was originally employed to study PS in studies belonging to the Gestalt school (eg Durkin, 1937; Dunker, 1945). It has also been used in many studies with weak theoretical associations of which ten relevant examples were quoted in Section 2.2.3 above. The technique is now most strongly associated with work in the information processing paradigm including in particular that of Newell and Simon (1972) with crypt-arithmetic problems. A number of information processing studies using protocol analysis in the field of science education were reviewed in Section 2.2.4.

Although protocol analysis offers the possibility of obtaining information that is not available in any other way, the method is not without its difficulties. Krutetskii (1976) amongst others warns of three problems in particular. Firstly, the student may be unsettled by the presence of the investigator and/or the microphone and by the fact of being asked to think aloud. Mujib (1980) regarded this as a serious difficulty. However, experimental investigation suggests that verbalisation of this sort does not influence PS performance (Flaherty, 1975). If this may be taken to indicate that PS processes are not affected, then this work supports the use of protocol analysis. Secondly, there is the danger that students may attempt to explain the solution (as they are often asked to in the classroom situation) rather than to think aloud. They may also attempt to rationalise or idealise their solution and to suppress thoughts which they may feel to be incorrect or irrelevant. Most reports suggest, however, that given careful instruction and a little training and/or occasional prompting this difficulty is quite easily overcome. Finally, Krutetskii warns of the dangers of the students misunderstanding their instructions and attempting to self-observe and describe their own thought processes. He cautions that their inexpert interpretations may confuse the data, but again recent reports suggest that careful instructions and some initial prompting appear to overcome any such tendency quite easily. In general, experience with protocol analysis in the studies reviewed earlier in the chapter strongly suggests that, providing care is taken over the potential difficulties referred to, the method can yield useful and valid data.

The collection and transcription[1] of recorded protocols is obviously rather time-consuming and the analysis of the data presents serious problems. The classification of the responses obtained using predetermined checklists of categories has already been referred to (Hateley, 1979; Webb, 1979; Mandell, 1980). Other workers have relied on generating general descriptions. Where the amount of data is substantial, generating descriptive categories in this way can be a laborious and difficult process (Bloom and Broder, 1950). Two alternative, though no less laborious, methods of handling the data from PS protocols have been described by Newell and Simon (1972) and Bliss and Ogborn (1979) respectively. The former employ Problem Behaviour Graphs which depict successive "information states", while the latter uses a species of network analysis to interpret and describe the data. Both these approaches will be examined in more detail in Chapter 4. In conclusion, it should be noted that it is rarely possible for protocol analysis to be entirely objective, insofar as the ways in which data is organised tend to be influenced by the point of view of the researcher. However, this need not be seen as a disadvantage providing that any such point of view is made clear.

---

1. Although something may be lost in transcription it has generally been found impracticable to work directly with audio-tapes.

2.5 A simple typology of examination items

It was noted in Chapter 1 that the nature of the task is obviously an important variable in the study of PS behaviour, and O-level chemistry examination items were classified in Section 2.1.1, as having well-defined goals and as depending for their solution upon information stored in the candidate's memory as well as that given in the item. However, before going on to describe the detailed planning of the first phase of the study, it is necessary to consider the different types of examination item in more detail. Such items can be classified in a number of ways. Three "styles" of questions are customarily employed which are distinguished by the nature of the answer demanded:

(i) **Objective items**, generally multiple-choice, in which the candidate selects the best out of a limited number of given alternative solutions to a problem.

(ii) **Short answer or structured response questions** consist of a series of items relating to a particular theme and often of gradually increasing difficulty; they demand only short answers (generally a few words although up to one or two short sentences) or short numerical calculations.

(iii) **Essay type questions** require the candidate to compose answers in the form of continuous prose; they may also demand diagrams and longer numerical calculations and may be structured into several parts.

Items are also usually classified in examination specificiations or "blueprints", in terms of a list of chemical topics and a set of educational objectives. The latter are customarily based on the taxonomy published by Bloom and co-workers (Bloom, 1956) and might typically include, for example:

(i) knowledge
(ii) comprehension
(iii) application
(iv) higher abilities.

These labels refer to a hierarchy of educational objectives which particular items are intended to reflect. However, a practical difficulty exists inasmuch as different students may have different knowledge and different experience. Thus an item presenting a novel challenge to one (assessing say "application") may be very familiar to another (assessing at best "comprehension").

A number of studies have reported the application of factor analysis to science examination scores (Lewis, 1961, 1964, 1969a, b; Chokotho, 1975; L'Odiaga, 1977). These have failed to validate the Bloomian categories and the factors identified tend to relate to item "style" and to chemical topics. Factors relating to general intellectual ability, and in particular numerical or mathematical ability, are also prominent.

For the present study it was necessary to be able to limit the range of items selected. Item "style" provided a reasonable basis inasmuch as each style makes a different set of demands on students and has some factorial as well as face validity. Thus PS strategies such as guessing and elimination apply chiefly to objective items, whilst only essay type items make significant demands on communication skills. However, a further breakdown was needed and neither chemical topic nor level of educational objective seemed attractive. As a number of different schools were likely to be involved in the study, difficulties were anticipated in selecting particular topics, nor was it felt to be desirable in terms of the model of PS adopted to tap only one or two areas of LTM. The unsatisfactory nature, from the cognitive research point of view, of the Bloom levels has already been referred to. Inspection of items from O-level science examinations in the light of the PS model, led to a tentative distinction between those thought likely to demand algorithmic processing and those which appeared to demand the processing of propositional knowledge. In practice the former proved to be largely items involving mathematical manipulations. In view of (a) the well-known difficulty which many students seem to find with mathematical items, (b) the numerical ability factor associated with performance in chemistry examinations, and (c) the potentially ambiguous nature of the proposed algorithmic/propositional distinction, the following simple typology was tentatively adopted for the first phase of the study:

(i) item style - objective or short answer or essay type

(ii) item content - mathematical or non-mathematical.

The usefulness of this classification was subsequently confirmed during pilot trials which are fully reported in the next chapter.

CHAPTER 3

THE PLANNING AND ADMINISTRATION OF THE FIRST PHASE OF THE STUDY

This chapter deals first with some preliminary planning details including the decision to undertake small-scale pilot trials. The conduct and findings of these trials are then discussed and the chapter concludes with a detailed account of the planning and administration of the first phase of the main study.

3.1 Preliminary planning

Early decisions were made regarding the methods to be adopted and the nature and scale of the study, including the types of item to be employed and the need for pilot trials. These decisions are reported below.

3.1.1 The adoption of protocol analysis. The aim of the first phase of the study was to obtain first-hand information about the way in which candidates tackle GCE O-level chemistry items. It was hoped to identify and describe the sorts of strategies adopted and the various ways in which such strategies break down when candidates fail to arrive at a correct solution. It was also hoped to identify and describe the ways in which the PS behaviour of successful candidates differs from that of the less successful. It was pointed out in the previous chapter (Section 2.4) that, despite difficulties in the interpretation of the data, protocol analysis is now widely regarded as the method of choice for the investigation of PS processes; indeed, in the context of this study, no other method offered any realistic prospect of obtaining the detailed information sought.

3.1.2 The scope of the study and the selection of items. The aims outlined in the preceding section had implications for the scope of the study. The intention was to develop empirically supported descriptions of PS with some claim to generality, and to make comparisons between the cognitive behaviour of more and less successful candidates. This implied the employment of a reasonable number and range of both test items and students. The model developed in the last chapter suggests that PS behaviour will reflect aspects of LTM organisation and these in turn will reflect learning experiences which are likely to differ at least in detail from one school to another. For this reason it was decided that the general validity of the study would be improved by using students from a number of different schools.

The study was explicitly concerned with examination behaviour and the selection of GCE O-level chemistry has already been referred to. It reflected, on the one hand, the author's background in chemistry and, on the other, the importance of this examination at the end of compulsory schooling. The Joint Matriculation Board's (JMB) examination was chosen because it was the only one in use at a reasonable number of schools in the region accessible to the author. The decision to use the GCE rather than the CSE examination was dictated by the small number of schools and candidates involved in the CSE in the area in which the work was undertaken. This reflected on the one hand trials of a proposed new (16+) examination in a number of schools and the fact that many others had only recently been converted from grammar to comprehensive schools and were not yet using the CSE with significant numbers of candidates.

The detailed selection of items at different stages in the study will be discussed at the appropriate points. However, an early decision was taken to exclude essay-type questions, at least from the first phase of the study. Consideration of 30 such items used in JMB examinations between 1974 and 1978 suggested strongly that they were not well adapted to covering a good range of topics and skills in a limited time. Their use would therefore not have been compatible with the aims of an exploratory study, particularly as it was anticipated that schools would feel able to allow only very limited time for work with their examination classes.

3.1.3 **The need for pilot trials.** The pilot trials were seen as an essential part of an iterative process in planning the first phase of the study. It was felt that a "dry run" would be helpful in reaching numerous detailed decisions regarding the conduct of the main study. In particular, trials would provide an opportunity to evaluate protocol analysis as a technique and to gain experience in using it. In the same way it might provide information regarding the characteristics of both the students and the examination items most likely to yield useful results.

3.2 Pilot trials

The reasons for holding pilot trials have already been mentioned. However, although the conduct of the trials and the analysis of the results took several weeks and the reports covered many closely written pages, these events were soon overtaken and superseded by the main study. Technical details will therefore be reserved for later sections and this part of the report will attempt only to provide a rather general picture.

The trials took place in two stages. The first stage was arranged informally when one lower-sixth-form chemistry student known to the author volunteered to record a PS protocol. This took place in November 1978, some six months after he had passed his O-level examination. The experience gained in collecting and analysing this, contributed to the planning of the second stage which involved six fifth-form chemistry students from a nearby school. This took place in December 1978, some five months before their O-level examinations. The teacher concerned was asked to select two of his best chemistry students, two of average ability and two of the least able students. He was also asked to avoid selecting any particularly shy or inarticulate student who might be expected to "dry up" in front of a tape-recorder.

3.2.1 **Test construction.** For the first stage, with only one student, a test was constructed using two short answer questions (comprising seven items in all) and eight objective items. Mathematical and non-mathematical problems were represented in both styles of item and the coverage of chemistry topics was fairly wide. All the items were original but were closely modelled on those set in JMB O-level examinations between 1974 and 1978. They were subject to independent expert scrutiny and correction before the trials.

Following satisfactory experience with this test, two parallel versions of similar length and general structure were compiled for the second stage. Some items from the original test were used again but a number of new items were also introduced. The two parallel versions had half their items in common to enable some comparison to be made of the PS behaviour of all six students. Half of the items, however, were unique to each version to enable a wider range of PS situations to be explored. The results of both stages of the trials will be discussed together in the sections which follow.

3.2.2 **The administration of the test and the collection of protocols.** After a little general conversation to establish a relaxed atmosphere, a typewritten set of instructions was discussed with each student before the test was started. These instructions referred to the purpose of the research and asked the students to "think aloud" as they attempted to answer the questions. Otherwise they were to proceed exactly as they would in an examination, writing down the answers and showing any "workings". They were also cautioned to avoid rationalisation and encouraged not to suppress apparently trivial or irrelevant thoughts. Three students (one good, one average and one poor as regards their usual performance in chemistry) tackled each version of the test during the second stage trials.

Although students were asked if they would like to do an unrecorded "dummy run" (a separate short-answer question had been prepared for this purpose) only two out of the seven elected to do so. Some prompting and encouragement was needed by most students for the first few minutes but after that all talked surprisingly freely. Where it was not apparent how a particular result had been arrived at, the author intervened to try to obtain more detailed information. Subsequent analysis of the tapes, however, indicated how difficult it was to avoid apparently leading questions or providing inadvertent cues; even the tone of an apparently neutral enquiry was sometimes taken by the student to indicate that he was or he was not on the right track and thus may have influenced his next step. It was clearly desirable to keep such interventions to a minimum. However the overall response of the students was very encouraging. None appeared to find the exercise an ordeal and the majority said that they had enjoyed it. Several voluntarily went on beyond the 30 minutes planned for the exercise and two independently expressed the opinion that the experience would help them in tackling examinations in the future. Interestingly, Gagné and Smith (1962) report that verbalising in this way may indeed have a facilitating effect on subsequent PS performance!

3.2.3 **The analysis of the results.** The seven students made 99 PS attempts spread over 31 items in approximately 230 minutes. Thus on average three attempts were made at each item and each attempt took a little over two minutes. Marking the students' written answers established that the rank order of their performances was identical to that provided by the teacher, with the single sixth-former heading the list. This gave at least mild encouragement to the hope that the results obtained were not unduly distorted by the need to think aloud and tended to confirm Flaherty's (1975) finding to this effect. The marks (ie proportion of successful attempts) ranged from 7% to 87% and averaged 46%, suggesting that the test as a whole was of roughly appropriate difficulty.

A few days' work quickly established that, as expected, it would be essential to prepare written transcripts. It was also clear that this laborious task could not be passed over to an audio-typist. The principal reasons for this were the technical nature of the language being used and the fact that many of the students' voices tended to become quiet and occasionally indistinct when they were facing a difficulty. Even the author, who was not only very familiar with the subject matter but had been present while the tapes were being made, often had to replay such passages a number of times before they could be understood. However, although the task was tedious, there was no serious difficulty in preparing transcripts for analysis.

The conduct of the analysis was an iterative process[1] and, given the limited sample and the rather broad range of items and student ability,

---

1. An attempt was also made to investigate the more formal approaches to protocol analysis using both problem behaviour graphs (Simon and Newell, 1972) and network analysis (Bliss and Ogborn, 1979). However, on account of the very limited generality of the data it was decided to postpone evaluation of these methods until more substantial material was available from the main study.

it proved difficult to generate descriptions of PS behaviour that were both general and well defined. On the other hand the initial impression that the protocols contained an enormous amount of valuable information, albeit of a case study kind, was in no way diminished as the analysis proceeded. Equally, the value of the PS model adopted was quickly confirmed as a useful framework for describing students' attempts. It was also soon established that the distinctions between short-answer and objective items and between mathematical and non-mathematical ones was a very useful one. The presence of a set of alternatives in the objective items clearly influenced the approach of all students, while mathematical items not only tended to elicit distinctive processing but were not usually seriously attempted at all by the lower scoring students. In giving a short account of the analysis it will be convenient to discuss short-answer, non-mathematical items as providing the model, and then to draw attention to the differences observed with the other types of item.

3.2.4 **A summary of the analysis relating to short-answer non-mathematical items.** Before summarising the results, the flavour of the data and of the analysis may be conveyed by a brief account of students' attempts at one item. An earlier part of the question had given $COCl_2(g)$ as the molecular formula of carbonyl chloride and they were told:

> "Using lines to represent covalent bonds, give the probable structural formula of carbonyl chloride."

All the seven students attempted this item and their protocols were analysed in terms of the PS stages identified in Table 2.3 which may be summarised as reads, understands, selects, operates and checks. There is little to say about **reading** the item, but when it came to **understanding** the goal (drawing a structural formula) one of the least able students was quite unable to proceed. Two of the students, while understanding what was wanted, seemed unable to **select** any way forward. Another selected the obvious strategy of combining valencies but generated incorrect numbers of bonds (C = 2, O = 1) and became confused in trying to combine this misinformation. The three most able students all succeeded. Two simply recalled all the valencies (propositional knowledge) and combined them correctly, whilst the third generated the valencies of oxygen and chlorine from propositional information about electronic configurations and the periodic table, only that of carbon being directly recalled. All the successful problem solvers put carbon in the centre from the start (often apparently intuitively, although in one case reference was made to recall of organic structural formulae) and seemed to use trial-and-error to sort out the bonds from there. The **operate** stage was thus successfully completed by three out of the four who reached it. The one who failed illustrated two sources of error. First he failed to recall essential information correctly, then even the incorrect knowledge he was using became confused at the combining stage. Thus carbon was said to have two bonds, and oxygen one, but he drew carbon with one and oxygen with two! This might speculatively be associated with inadequate rehearsal in STM or even with STM overload.

The protocols relating to every item were, of course, analysed in considerably more detail than has been indicated above. However, insofar as the results of the pilot trials were superseded by those of the main study it will be sufficient to summarise here the main conclusions. It will be convenient to refer again to each of the five PS stages.

(i) **Reads** - The close relationship between reading and understanding was apparent from the fragmented way in which students often read items but otherwise little of interest was recorded.

(ii) **Understands** - This relates to the students' analysis, reorganisation, translation and encoding/recognition of what he reads in relation to his LTM store and is closely associated with the **select** stage. There were only two occasions on which failure to understand the goal or a critical idea blocked any attempt; however, it seemed probable that limitations in understanding often interfered with subsequent processing, for example as regards the cueing of appropriate parts of LTM store.

(iii) **Selects** - It was difficult to separate the selection of the starting point from the selection of the method of attack or strategy. The following strategies were identified:

(a) Direct recall of an answer through location of appropriate associations in LTM.

(b) Location of relevant general propositional knowledge in LTM (including generalisation from specific equivalent cases) and rational combination of this to generate a response.

(c) "Serial processing": that is, unsystematic processing of given information in the general direction of the goal. It might be regarded as a failed version of the strategies described in (a) or (b) based on inadequate characterisation (understanding) of given information or of the goal.

(d) Generation and evaluation of hypotheses (though rarely demanded and adequately tackled by only a few of the most able students).

Errors associated with this stage were failure to locate a strategy at all (although the item was understood) and the adoption of "serial processing" which led to a variety of inadequate responses ranging from mere paraphrases of the information given, often with reference to some stereotype, to more directed efforts which still fell short of reaching the required goal. No errors were recorded with regard to the combining of propositional knowledge or to the evaluation of hypotheses.

(iv) **Operates** - This stage follows directly from the strategy selected and there is little to say about its successful application. However, two types of error were identified relating to the operation of LTM, and two relating to the operation of STM/WM.

(a) Recall failure - failure to recall essential specific or general information from LTM (attributed to its absence or to failure to locate/cue it);

(b) Incorrect recall - may again be specific or general (attributed to wrong/inadequate learning or miscuing);

(c) Failure to keep-in-mind and use essential information given (although it had been read and understood);

(d) Failure to keep accurately in mind information previously generated in the course of a solution (both this and the previous error were attributed to attention failure, probably associated with the overloading of STM; they would normally be referred to as "careless" errors).

(v) **Checks** - There was no obvious direct evidence of (a) checking the answer against the goal (although in most solutions this was almost a logical necessity and could be inferred), (b) checking for consistency with the student's existing cognitive map or (c) checking of the processing which led to the answer.

Although the above summary does not fully reflect the conclusions and speculations generated by the pilot trials it will suffice at this stage. The issues raised will be more adequately explored in discussing the results of the main study. It remains to note briefly some of the special characteristics associated with students' attempts at mathematical and objective items.

3.2.5 **Special characteristics relating to mathematical items.** All but the most able students tended to express negative attitudes towards mathematical items. These were often reflected in explicit remarks about disliking such items, and on several occasions students stated that they would "skip" calculations in an examination, leaving them to be attempted at the end if time allowed. Otherwise the only significant new information related to the **select**, **operate** and **check** stages.

(i) **Selects** - In addition to the recall of specific and general propositional knowledge, the recall of algorithms was prominent. Where the student experienced difficulty two approaches were noted in addition to "serial processing". These were labelled "naive assumption" and "naive generalisation". The former referred to a tendency to seek and exploit purely mathematical relationships within the data without reference to chemical principles, and the latter to a tendency to generalise from recall of a superficially similar case without reference to relevant differences in the chemical system.

(ii) **Operates** - In addition to the errors reported for non-mathematical tasks, "careless" errors were observed in the application of mathematical algorithms and in performing elementary arithmetical operations.

(iii) **Checks** - One case was observed of explicit checks of all the three types discussed.

3.2.6 **Special characteristics relating to objective items.** Features of interest were again largely confined to the **select** and **operate** stages. It was interesting to observe that students usually concentrated on the stem, generating related propositional knowledge from LTM and combining and matching this with the multiple choice responses, much use being made of elimination. The role of cuing by the given responses was unclear. Correct answers generated from data in the stem were sometimes rejected through failure to identify them with the key, and occasionally through unwillingness to select certain responses - for example "They never use none-of-the-above"; "Can't be A, the last two were As"! Guessing was fairly frequent, often after the elimination of certain responses. A tendency was tentatively identified to guess the least familiar remaining response when the stem was well understood, and the most familiar when the student was completely in the dark. An additional source of error was the occasional misuse of the coding in multiple completion items.

3.2.7 **Postscript.** The pilot trials influenced the planning and conduct of the main study in many ways both obvious and subtle but it is doubtful if any useful purpose would be served by attempting to list them all. Speaking generally, it provided both valuable experience, particularly in the use of protocol analysis, and strong encouragement to proceed. Where experience in the pilot trials exerted a specific influence on subsequent decisions this will be indicated in the account which follows.

3.3 The planning and administration of the main study (phase 1)

Following the pilot trials in November and December 1978, detailed preparations were made to carry out the first phase of the main study during March and April 1979 immediately before students sat for their O-level examinations. It was felt that longer recordings covering a representative range of items in each of the four main classes (short-answer or objective items with mathematical or non-mathematical subject matter) and using larger numbers of students spread over several schools, would lead to clearer and more generalisable insights into PS strategies and errors. As a result of the trials it was decided to direct attention first to the short-answer non-mathematical items. These appeared to elicit a range of basic strategies and errors common to PS attempts at all items. The account which follows deals with the further analysis of such items and with the subsequent construction and administration of a short-answer non-mathematical test to twenty students from four different schools. The collection of 60-minute PS protocols from each student together with certain supplementary information is also fully described.

Two similar tests were prepared and administered covering, respectively, short-answer mathematical questions, and objective items of both mathematical and non-mathematical types. A further forty students contributed approximately 40 hours of protocols. In the event, a decision taken following the analysis of data from the first test led to a decision to hold over these tapes for study at a future date. This decision is documented at the beginning of Chapter 5. Meanwhile the account will focus solely on short-answer non-mathematical PS.

3.3.1 **An analysis of short-answer non-mathematical questions.** It was considered essential to make sure that the test contained a representative sample of short-answer non-mathematical items. To facilitate this, a basis was sought for the analysis of past JMB examinations. It was speculated that the single feature of an item most likely to affect students' PS behaviour was the nature of the goal to be achieved. Limited support for this idea was obtained from the pilot test data. For example, the generation and testing of hypotheses was associated only with two items asking for explanations of phenomena. No alternative basis appeared promising and the results of an anlysis of the information demanded in the items appearing on five JMB O-level chemistry examinations between 1974 and 1978[1] is presented in Table 3.1.

The six categories of goals (which have also been expressed in terms of the tasks demanded) should be self-explanatory and short descriptions have been given of all the examples found of each. In the course of the analysis, and on the basis of the pilot trials, an estimate was made of the range of strategies likely to be employed by students tackling each group of items described. Inspection of the table will show that this estimate helps to confirm the expectation that attempts based on the categories identified would also be representative in terms of the PS behaviour elicited. The construction of the test is described in the next section.

---

1. The following papers were analysed: June 1974, November 1974, June 1975, June 1977 and June 1978. The "missing" papers were out of print at the time.

Table 3.1 Analysis of goals/tasks demanded in short-answer items in five JMB O-level chemistry examinations (with examples)

Goal/task demanded by item	Expected basic strategy *	Number of items	
1. Definition/defining	S1		3
2. Class identity/classifying			9
(a) changes as exo/endo-thermic (no evidence)	S1	2	
(b) specie(s) as acid(s) given equation as data	S1/S2	1	
(c) name reducing agent(s) from list of 3 given	S1	1	
(d) identifying sets of data as examplars of concepts	S2	5	
3. Specific name/naming			24
(a) chemical substance as product of reaction	S1/S2	17	
(b) chemical substance (on other basis)	S1	4	
(c) specific process	S1	3	
4. Statement or description/stating or describing			21
(a) use of chemical substance	S1	1	
(b) reaction for given interconversion	S1/S2	1	
(c) experimental procedure	S1/S2	1	
(d) standard tests	S1/S2	9	
(e) non-standard tests	S2/S3	2	
(f) "What would happen/be observed if ...?"	S1/S2/S3	7	
5. Explanation/explaining			12
(a) familiar qualitative phenomena	S1/S2	4	
(b) familiar quantitative phenomena	S1/S2	1	
(c) discrepant phenomena	S2/S3	2	
(d) comparative phenomena	S1/S2	1	
(e) experimental procedures	S1/S2/S3	4	
6. Formula/giving formula			6
(a) molecular formula of familiar compound	S1	1	
(b) structural formula of familiar compound	S1/S2	3	
(c) structural formula given molecular formula	S2	1	
(d) other (novel)	S2	1	
	TOTAL		75

* Key: S1 = simple recall; S2 = recall and combination of propositional information; S3 = generation and testing of hypotheses.

3.3.2 **The construction of the test.** The length of the test was determined by two factors: (i) the need to ensure that a representative range of tasks was included and (ii) the time likely to be available for working with the students.

As regards tasks, all the categories identified in the analysis of past JMB papers were employed, namely:

Task 1 Defining - defining a given concept or stating a given principle;

Task 2 Classifying - attaching a concept label to a given example or vice versa (the example may be simply named or relevant data may be provided);

Task 3 Naming - giving the **specific** name of (usually) a chemical species as a product, reagent, constituent, etc;

Task 4 Stating or describing - giving statements/descriptions regarding:

4.1 processes or procedures ("How would you ...?", "What would you do ...?" etc);

4.2 occurrences ("What would happen/be observed if ....?" etc);

Task 5 Explaining - giving explanations regarding:

5.1 processes or procedures;

5.2 occurrences;

Task 6 Giving formulae - giving:

6.1 molecular formulae;

6.2 structural formulae.

Tasks 2 to 5 in particular each cover quite a wide range, and further analysis suggested that a minimum of 20 to 25 items would be needed to compile a reasonably representative test. In the event, schools were prepared to allow one double period amounting to about 75 minutes for work with each participating student (see Section 3.3.3 below). During the pilot trials most students averaged about 2 minutes per item and in view of the proportionately smaller loss of time in preliminaries during a longer session, it was felt that a test of about 40 items would not be unreasonable.

Nine short-answer non-mathematical items were composed in line with the JMB syllabus and format and comprising 46 items/tasks. Two of these were questions successfully employed in the pilot trials, the remainder being newly written. All were modelled on questions from past papers and were subject to independent expert scrutiny before being finalised. Table 4.3 in the next chapter shows the distribution of items against tasks. It must be emphasised, however, that the present task-analysis was merely a tool to facilitate the compilation of a representative test and that no other significance should be attached to it.

The test was a little longer than originally intended, tasks 2 and 3 (classifying and naming) being somewhat over-representated relative to other areas. This might have been corrected at the expense of making some of the questions rather stilted, but it was felt that 46 items would be within the scope of at least some of the students and this proved to be the case. It also allowed some leeway in case a student encountered items which he was unable to attempt by providing a somewhat wider range of alternatives. To ensure as even a coverage of the tasks as possible for pupils who worked more slowly, questions were ordered so as to distribute examples of each task equally between first and second halves of the test.

As regards the subject matter covered, the participating students were in the latter part of their second term in form V and had already done their "mock O-levels". It was therefore assumed that the test could be based on the full syllabus. Although it was clearly too short for any systematic attempt to cover every topic, a reasonably wide and representative range was achieved. It was considered (parenthetically) of interest to include one topic in a variety of contexts and this was done with oxidation in four separate items. It was hoped that this might throw light on the influence of context and the set of items is briefly described in a footnote below.[1]

The test was printed on A4 paper, spaces being provided for students to write their answers on the question paper in the usual way. A front cover was added including the standard instructions used by the JMB. A copy of the test appears in Appendix A.

3.3.3 **The test population.** Since the study was concerned with the PS behaviour of examination candidates, it was considered essential to ensure that the examination situation was reflected as closely as possible and the timing of the study a few weeks before the O-levels has already been mentioned. It was anticipated that students would be fully geared up for examinations and they were encouraged to regard their participation as examination practice.

The pilot trials had drawn attention to the time needed for the transcription and analysis of protocols and it was decided that about 20 hours of recording would be an acceptable target. Although most reports do not refer to giving students any prior training in the "thinking aloud" technique and satisfactory results were obtained without this in the trials, others, particularly Krutetskii (1976), have argued the value of such training. However, schools were understandably cautious about the amount of time for which they were willing to release students during their final examination preparation. As already mentioned, it was possible to arrange only one double period (about 75 minutes) with each student and it was thus decided to proceed on the basis of 20 1-hour tapes without training.

---

1. Item 3d asks which of three previously named gases normally acts as an oxidising agent, no evidence being provided. Item 6c asks for identification of the oxidising agent in a verbally described displacement reaction from which evidence could be deduced. Finally, 7a asks for a definition of oxidation in terms of changes at the atomic level and 7b provides an opportunity for directly applying this to given ionic equations.

To eliminate systematic effects due to a particular school environment, four different schools were each asked to provide five students for the study. As in the trials, it was suggested that particularly shy or inarticulate students should not be chosen and that volunteers would be preferred. In addition it was stipulated that a range of achievements should be covered from the very able down to students who were only borderline candidates. The pilot trials had suggested that students performing much below a pass grade might contribute little useful information. In any case, students unlikely to pass chemistry generally study other options and would not have been available in most schools.

Although gender was not regarded as a major variable in this study, it had been hoped to test equal numbers of boys and girls. Unfortunately this proved impossible when one of the girls' schools originally approached was unable to participate at the last minute and only a boys' school could be found to deputise at short notice. There were thus 15 boys and 5 girls in the sample.

As some schools wished to be anonymous, each was assigned a letter of the alphabet (C, D, E and F) for reference purposes. The anonymity of individual students was also protected, only their first names being recorded, and each being assigned a three digit reference number (008-028). School C provided 6 students instead of the 5 requested and the extra protocol was used when one student from school D (015) was not present during testing.

3.3.4 **The administration of the test and the collection of the protocols.** At each school, arrangements were made to meet the students participating in the study as a group before they tackled the test. A set of specially written instructions regarding the collection of the protocols (see Appendix A) was distributed, and the purpose of the study and the role of the students were discussed. Emphasis was placed upon the importance of trying to think aloud as opposed to making post-hoc rationalisations and an example was discussed to clarify these instructions. Students were also told to write any rough working on the paper and to write out their answers in full. To encourage this, and a serious attitude to the exercise, they were told that the papers would be marked although this was not over-emphasised. While one purpose of meeting the students was to shorten the time needed for giving individual instructions later, an equally important aim was to "break the ice" at a personal level in the hope that they would settle down more quickly when it came to the individual sessions.

Following the preliminary meeting, each student in turn spent about an hour working on the test individually, thinking aloud as he did so while a tape-recording was made of what he said. A few minutes before the start was spent in an informal chat with the author to ensure that the instructions were understood and to try to put the student at ease. A "record of protocol collection" form was completed (see Appendix A) at the same time. On this the student's first name and school were recorded and an index number, copied onto the tape cassette, was assigned. Five general questions were asked in an attempt to establish a rapport with the students. These related to interest in science and chemistry, the perceived difficulty of these subjects, A-level and career plans and interests and hobbies. As many of the replies were indecisive,it was not possible to encode them as additional data.[1]

Students generally seemed relaxed by the time these preliminaries were completed and every effort was made to achieve this. They were told

---

1. It would have been inimical to the purposes of the exercise to have pursued these questions further and, while the data might have been useful, it was not related to the main purpose of the study.

that the first question would be treated as something of a trial run and were then asked to start. A certain amount of prompting and encouragement was used to get students to vocalise their thoughts during the first few minutes but in most cases they proceeded quite satisfactorily after the first two or three items. The author's contribution quickly became limited to reminders to keep verbalising their thoughts, to cautioning against post-hoc rationalisation ("Was that **really** how you got that answer?") and to providing general encouragement while trying to avoid feedback. It very quickly became clear that encouragement alone, regarding the success of attempts at verbalising and in achieving acceptable answers to the questions, was much the most potent way of achieving a satisfactory response.

Recordings were made using a conventional cassette tape recorder and standard C60 tapes. Actual recording time varied from about 45 minutes to over 60 minutes and the number of items attempted from 32 to 46 with a mean of over 38. Table 4.3 in Chapter 4 shows that practically all students answered the first seven out of the nine questions on the paper and that an ample number of responses was obtained to items representing each of the tasks identified.

3.3.5 **The collection of supplementary data.** Two sets of supplementary data were collected to allow students to be assessed and compared relative to an independent standard. These were scores on a test of general intelligence (AH4) and "mock O-level" marks in chemistry. The latter were obtained from their teachers and the AH4 test was administered a few days after the recording session in most schools, but immediately beforehand at school E. Correlations between science achievement and measures of general intelligence have been widely reported and the possible relevance of this variable to the study was noted in Chapter 1 (Section 1.3.2 (ii)). The "mock O-level" scores were intended to provide a basis for comparing students' performance during the study with that under more normal testing conditions and their actual O-level grades were also obtained when these became available later. The additional data collected was added to the bottom of each student's "record of protocol collection" form in the space provided for the purpose, together with the score obtained by marking his test script (see Section 4.1). The analysis and interpretation of the data is described in the next chapter.

CHAPTER 4

THE RESULTS OF THE FIRST PHASE OF THE STUDY

Basic statistical information relating to the test and to the other quantitative data collected is reported in the first section of the chapter. Alternative methods of analysing the students' protocols are then evaluated and the employment of a simple category analysis for this purpose is described. A detailed account of students' PS strategies and errors follows and evidence regarding the checking of solutions is also considered. The chapter concludes with a discussion of some of the apparent interrelationships between students, strategies, errors and test items.

4.1 The statistical characteristics of the test

Twenty students worked on the test for times ranging from 45 to 62 minutes and attempted between 32 and 46 items each. Normative statistics were calculated to establish that the characteristics of the test matched those which would be anticipated under ordinary examination conditions. The students' written scripts were marked in the usual way, one mark being allotted for each correctly answered item. For tasks 1 (defining), 4 (stating/describing) and 5 (explaining), half marks were awarded for partially correct answers. Circumstances at the schools limited the time which some students were able to spend on the tests and restricted the number of items attempted. All scores (mark totals) were therefore expressed as a percentage of the number of items attempted. In view of the fact that the majority of students normally complete their O-level examination papers within the time allowed (ie that speed of working is probably not in itself a major factor in examination performance) this appeared to be a legitimate way of obtaining an overall measure of performance on the test.

The raw data, together with the mock O-level marks supplied by teachers and the students' scores on the AH4 test of general intelligence, are presented in Table A1 in Appendix A; students' subsequent O-level grades were obtained from the schools the following September and have been added to the table. Some analyses of these data are discussed in the following paragraphs.

4.1.1 **Overall test statistics.** General statistics relating to the tests and to other data collected are summarised in Table 4.1 below.

Table 4.1 General statistics relating to the quantitative data collected*

	Score range	Mean score	Standard deviation	Reliability KR20
Test %	35 - 87	57.4	17.4	0.87**
Mock O-level (%)	39 - 92	-	-	-
AH4 total score	75 - 115	98.9	11.7	-

* As schools each set their own mock O-level examinations no overall statistics could be calculated. The AH4 is a test of established reliability so no attempt was made to estimate this.

** The Kuder-Richardson reliability was calculated using the formula

$$r = \frac{n}{n-1}\left(\frac{\sigma^2 - \Sigma\, p(p-1)}{\sigma^2}\right)$$

where n was the number of items (46), σ was the standard deviation of the scores (rescaled for a 46-item test), and p was the ratio of correct answers to **attempts** at each item; p was calculated in this way to avoid over estimating the reliability.

The reliability (Kuder-Richardson formula 20) of the test was of a satisfactory level and the mean and standard deviation acceptable considering that all the students were expected by their teachers to pass O-level. The range of scores (35% - 87%) was exactly what would have been expected for a group including borderline cases and high performers and was closely similar to that of the mock O-level marks awarded in the different schools (39% - 92%).

4.1.2 **The validity of the test.** The validity of a test refers to the extent to which it measures what it is intended to measure. In the present case the ultimate focus of attention was the PS behaviour of GCE O-level chemistry candidates and there was no direct way of assessing the validity of the information collected in the protocols. However, the validity of the test as the measure of appropriate performance could be judged. Its face validity (or content validity) was assured by the way in which it was constructed; that is to say it was compiled from a representative range of tasks very closely modelled on those found in the actual (criterion) examinations. Despite this, it remained possible that the unusual circumstances in which the test was administered might have influenced students' performances. To assess this, the concurrent validity of the tests could be judged by calculating the correlation of students' scores on the test with those obtained on a criterion measure. At the time of the study no true criterion measure was available but the schools' own mock O-level examinations were regarded as a reasonable approximation. Table 4.2 below shows that a very satisfactory correlation ($r = 0.87$, $p < 0.01$)[1] was obtained. It was therefore assumed that the test was a valid measure of students' examination performance. This was subsequently confirmed when the students' actual O-level examination grades became available some months later and were found to correlate satisfactorily with the test scores ($r = 0.70$, $p < 0.01$).[2] In the absence of any evidence to the contrary it was judged likely that, given both face validity (appropriate tasks) and concurrent validity (appropriate performance), the test was likely to be eliciting PS

---

1. Because schools used different mock examinations, separate product moment correlations were calculated for each school and averaged using Fisher's z transformation to obtain the best estimate of the population correlation (Lewis, 1973).

2. The students participating in the study all passed with a grade A or B; a biserial correlation was therefore calculated (Lewis, 1973).

behaviour similar to that elicited in an O-level examination.

Table 4.2 Intercorrelations of test, mock O-level and AH4 scores

	Test	Mock	AH4
Test (%)	1.00	-	-
Mock O-level (%)	.87*	1.00	-
AH4 total score	- .04	.02	1.00

* significance $p < 0.01$

4.1.3 **The relationship of performance to general intelligence.** The data presented in Table 4.2 above show that there was no significant correlation between either the test or the mock O-level scores, and students' general intelligence as measured by the AH4 test.[1]

This suggests that, although all three scores showed a reasonable spread (Table 4.1), "intelligence" is **not** a factor influencing performance in chemistry within the highly selected population involved in the study.

4.1.4 **Item analysis.** An analysis of the performance of the individual test items was carried out to pinpoint any particular anomalies that might need to be considered before analysing the protocols. Table 4.3 shows the indices of facility (p) and discrimination (d)[2] calculated, and also the number of students (N) attempting each item. For convenience during the construction of the test and the analysis of results, items are identified in the table with respect to the task they represent as well as by their item numbers in the test.

In general, although some items were obviously too easy ($p > 0.80$ for six

1. Separate correlations with the different AH4 sub-scores (verbal/numerical and spatial) were equally low and non-significant.
2. Discrimination indices, d, were was calculated using "top" and "bottom" groups of 6 for items attempted by 12 or more students and of N/2 or (N - 1)/2 when $N < 12$.

Table 4.3 List of tasks, items and item statistics

Task	Task/item number		Facility index p	Discrimination index d	N
1. Defining	1.1	9d(ii)	.93	.17	7
	1.2	7a	.70	.50	20
	1.3	4a	.59	.50	11
2. Classifying	2.1	3d	.60	.67	20
	2.2	6c	.50	.00	18
	2.3	7b(i)	.65	.67	20
	2.4	7b(ii)	.65	.83	20
	2.5	9d(ii)	.67	.67	6
	2.6	5a	.80	.17	20
	2.7	5b	.65	.67	20
	2.8	5c	.65	.33	20
	2.9	9a	.73	.20	11
	2.10	9c(ii)	.71	.17	7
3. Naming	3.1	7c(i)	.40	.67	20
	3.2	7c(ii)	.60	.67	19
	3.3	8a	.71	.58	14
	3.4	9c(i)	.72	-.13	9
	3.5	1a(i)	.85	.50	20
	3.6	1a(ii)	.85	.50	20
	3.7	3a(i)	.60	.67	20
	3.8	3b(i)	.90	.17	20
	3.9	3c(i)	.85	.33	20
	3.10	6b	.72	.33	18
4. Stating/describing: 4.1 procedure/ process	4.1.1	3a(ii)	.35	.67	20
	4.1.2	3b(ii)	.55	.17	20
	4.1.3	3c(ii)	.60	.50	20
	4.1.4	3b(i)	.50	.33	20
	4.1.5	8c	.38	.25	12
	4.1.6	6d(ii)	.24	.33	17
	4.1.7	6d(ii)	.00	.00	17
4.2 occurrence	4.2.1	1c(i)	.10	.33	20
	4.2.2	1c(ii)	.25	.67	20
	4.2.3	4d(i)	.50	.20	10
	4.2.4	4d(ii)	.50	.60	10
5. Explaining: 5.1 procedures	5.1.1	4b	.65	-.50	10
	5.1.2	6a	.92	.08	18
5.2 phenomena	5.2.1	8b	.50	.67	14
	5.2.2	4c	.65	.70	10
	5.2.3	1d	.75	.67	20
	5.2.4	1b(i)	.35	.50	20
	5.2.5	1b(ii)	.45	.33	20
	5.2.6	6c	.24	.50	17
6. Giving formulae: 6.1 molecular	6.1.1	2a	.78	.58	20
	6.1.2	5d	.60	.70	20
6.2 structural	6.2.1	9b	.25	.17	12
	6.2.2	2b(ii)	.35	.83	20
				Total	767

items out of 46) this was to be expected for pupils of the high calibre involved and did little to detract from the interest of examining the strategies used in solving them. They might, indeed, be expected to provide an interesting contrast with more difficult items. All these easy items discriminated positively and some quite highly (p = 0.50 in two cases).

At the opposite end of the scale one item (task 4.1.7, item 6d(ii)) was eliminated. No student answered it correctly and it was obvious from the protocols that this was largely due to unsatisfactory wording which led to confusion with the related item preceding it. It was ignored in subsequent systematic analyses, and protocols relating to it were considered in conjunction with the related item (task 4.1.6, item 6d(i)). Only one other item (task 4.2.2, item 1c(ii)) had a facility less than 0.20; however it discriminated well (d = 0.33) and students' protocols indicated no obvious fault. The difficulty of the item was not felt to detract from the interest of analysing the way in which students tried to tackle it.

Owing to the small numbers of students involved, the values of the indices of discrimination could not be regarded as particularly reliable. However, it was considered desirable to re-examine the two items which showed negative d values. Task 3.4 (item 9c(i); d = -0.13) was attempted only nine times and revealed no obvious fault after rein-specting both the item and the associated protocols. The discrimination index of -0.50 on task 5.1.1 (item 4b) seemed, however, to reflect a somewhat more serious problem in that a number of students who had scored highly on the test as a whole failed to make an important distinction. However, both the item and the way it was marked seemed fully justified and the item was retained for subsequent stages of the analysis.

Thus one item (task 4.1.7; item 6d(ii)) was eliminated and the remainder of the chapter is devoted to the analysis of protocols relating to the remaining 45.

4.2 The processing of the protocols

The section which follows describes the transcription of the protocols and reports the theoretical and practical evaluation of alternative methods of handling the data concluding with an account of the analysis conducted.

4.2.1 **The transcription of the protocols.** The need to make repeated comparisons back and forth between different protocols and between different parts of the same protocol made it impracticable to work direct from the audio-tapes. It is, indeed, normal practice to make written transcriptions of PS protocols prior to analysis although a certain immediacy is inevitably lost in the process (Cowan, 1977). Owing to the technical nature of the subject matter and other factors external to the study, a handwritten transcription was made in the first instance. This proved a demanding and tedious task, each one-hour tape requiring on average six hours for accurate transcription. The written transcripts were subsequently typed and copies are given in Appendix A. A number of conventions were adopted to try to reflect as fully as possible the character of the protocol including hesitations, false starts, emphatic or enquiring tones and so on. This improved the immediacy of the protocols and, although not explicitly used in the analysis, sometimes assisted in their interpretation. A key to these conventions is provided in the appendix.

4.2.2 **General considerations regarding protocol analysis.** Before evaluating alternative approaches to the analysis of the protocols, it is necessary to consider the characteristics of the information which was wanted from them. The aim of this phase of the study was to find out how candidates tackled examination items, to describe the strategies they adopted and how these broke down, and to find out how the PS behaviour of the successful student differed from that of the less successful.

It seemed clear that the analysis should aim, so far as the data allowed, to generate descriptions of strategies which were (i) usefully general, (ii) unambiguous and (iii) of explanatory value. The identification of a limited number of strategies covering between them as wide a range of observed behaviour as possible, was a primary intention of the study. Indeed, a relatively large and representative range of items was employed in the test in the explicit hope of facilitating a usefully general description of PS behaviour. The virtue of such generality lies not only in parsimony for its own sake but in increasing the usefulness of the information as regards its implications for educational practice. The importance of avoiding ambiguity is obvious. Ideally, any descriptions of strategies should be such as to enable clear discriminations to be made on the basis of available evidence. Finally, strategies should be described so as to facilitate the interpretation of students' PS successes and failures with reference to an acceptable explanatory model. In particular, perhaps, they should provide a basis for identifying the underlying causes of students' errors.

Three different approaches to the analysis of protocols were considered. These were: (i) simple category analysis, (ii) problem behaviour graphing, and (iii) network analysis. Each was evaluated not only on the basis of its description in the literature but also through its trial application to protocols. All three are briefly described and evaluated below.

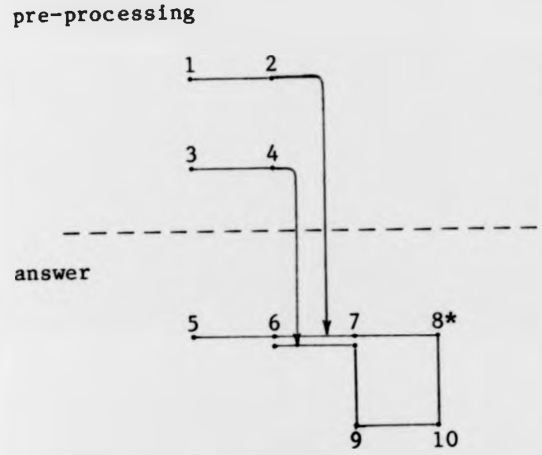
4.2.3 **Simple category analysis or iterative generalisation.** The phrase "iterative generalisation" has been coined here to describe what is probably the most common method of protocol analysis. In its simplest form it may amount to little more than a generalised description of data from a single case study. However, as the numbers of subjects and/or problems increase, initial tentative descriptions tend to be progressively modified and generalised to accommodate more and more of the data. As this process proceeds and descriptions of general categories are refined, earlier assignments of particular labels to particular cases may need to be reconsidered and the analysis proceeds in an iterative manner. The descriptions generated in this way may tend to reflect a particular view, theory or model of the PS process (Durkin, 1937; Dunker, 1945; Krutetskii, 1976; Larkin and Reif, 1979) although this is not necessarily the case (Bloom and Broder, 1950; Cowan, 1977).

The analysis of the protocols obtained in the pilot trials was conducted by the method described above which was also tried out on an item from the main study which was attempted by all 20 students. Although demanding and time-consuming, it led to useful if still tentative descriptions of PS processes at a high level of generality (eg direct recall, propositional reasoning, serial processing etc). As regards the unambiguous assignment of individual cases to particular descriptive classes, it is in the nature of the method that such classes are operationally defined by the examples assigned to them. Any ambiguity is thus confined to the way in which the protocol is interpreted, that is in deciding what to make of particular verbalisations. While many may be clear, others may involve a degree of interpretation and thus be less easy to classify objectively. It is, however, characteristic of protocol data that this level of interpretation is unavoidable whatever means of analysis is adopted.

Finally, in conjunction with the PS model developed in Chapter 2, the processes identified showed high potential in terms of their capacity to generate useful explanations of PS behaviour, including failure to solve items (see Section 3.2 relating to the pilot trials).

4.2.4 **Problem Behaviour Graphs.** "Problem Behaviour Graphs" (PBGs) were developed by Simon and Newell (1972) as a means of representing individual PS protocols. They are based on an information processing model of PS behaviour and were intended to facilitate the development of a "full theory of human problem-solving", and more particularly to lead to the construction of computer programmes that can model human PS behaviour. Simon and Newell claim that the method provides a means of encoding verbal protocols "with a relatively high degree of objectivity and reproducibility, and in very concrete terms, so that the content of the problem-solving steps (is) retained".

A PBG is constructed in a hypothetical "problem space" defined by those states of knowledge concerning the problem (including the givens and goal states) assumed to be accessible to the problem-solver through the application of a set of operations which he is able to perform on such knowledge states. The graph represents, in a formal manner, the sequence of information states which are reflected in the protocol and employs a number of conventions. For example, each information state is represented by a node (or sometimes a "box"), a succession of such states deriving from one another being linked by horizontal lines. A return to an earlier information state is indicated by placing a new node beneath the one in question and joining them with a vertical line. The nodes are numbered according to their sequence in the protocol and the infomation represented by each node is presumed to include knowledge of givens and of goals and of the attempted solution route up to that point, in addition to the specific information generated by processing. An example based on the attempt of student 020 on item 5a (and employing additional conventions generated in the course of evaluating this approach) is shown in Figure 4.1.



Key: 1 number of electrons (e^-)
2 equals number of protons (p^+)
3 mass number (M)
4 equals protons plus neutrons (n^o)
5 isotope (I)
6 is same number p^+; different number n^o
7 thus same number e^-; different M
8 thus X and Z (*answer)
9 I means different M
10 matches X and Z (check)

Figure 4.1 Problem Behaviour Graph (PBG) of student 020; item 5a, test 03

The student was given tabulated data regarding the mass numbers and the numbers of electrons in seven neutral atoms, and asked to identify which two were isotopes of the same element. The rather cryptic key should enable the solution to be followed. Nodes 1 - 2 and 3 - 4 represent information generated while the item was being read. Nodes 5 - 8 represent the main solution route. Nodes 6 and 7 each represent two separate pieces of information processed in parallel. Only one was used to generate the solution, however the other was subsequently used as a cross-check (nodes 9 - 10). The arrows from nodes 2 and 4 show that the information they represent was applied in generating node 7 from node 6.

Although the construction of PBGs offers a neat and convenient way of coding successive information states, it does not include all the information available from the protocols (regarding for example the processing between information states, the student's certainty or doubts regarding the progress of the solution, etc) and on its own offers little by way of interpretation. Two additional stages are needed. The first involves identifying the processes which are operating in the problem space. These may sometimes be available from the protocols but often have to be inferred. Secondly, by generalisation from the association of particular information states with particular processes, a "production system" is inferred. This specifies sets of "conditions" leading to particular mental "actions" on the part of the problem-solver and allows a "programme" to be written whose execution will simulate the PS behaviour of the subject. Interestingly, in recent work on the PS of university physics students, Larkin and Reif (1979) have described the generation of production systems from protocols without the intervention of PBGs.

Although this approach to the analysis of the protocols has some attractive features, its use has been restricted to case studies. Simon and Simon (1980) for example have generated very explicit production systems which reflect the behaviour of two individuals on a restricted set of physics problems. The PBG is a specific not a general description, indeed on the item used for the illustration in Figure 4.1

20 students produced 19 different graphs. Although generalisations could certainly be made by applying iterative methods to the graphs, or by the subsequent analysis of processes and production systems, this can be done direct from the protocols. It should also be noted that PBGs are constructed within a problem-space already selected by the student and focus attention on the detailed analysis of his subsequent behaviour. This leaves aside vital questions relating to the understanding of the problem and the selection of a strategy, and work in this paradigm has been criticised for focusing attention solely towards the processes of PS without adequate regard to these other important aspects (Klein and Weitzenfeld, 1976). As regards the apparent objectivity of the method, this relates largely to the construction of the PBG itself. Insofar as both the processes and the production systems are inferred, the method makes similar demands on the interpretive skills of the researcher as any other method of analysing protocol data. The solutions generated are essentially programmes whose implementation simulates the PS behaviour observed. These have descriptive rather than explanatory value and would demand further interpretation to meet the demands of the present study. Thus although the analysis imposes a disciplined structure on the data, in itself it contributes little as regards generalisation and explanation, and confines its attention largely to only one of the five PS stages identified. Though well adapted to case study work in the computer simulation paradigm, it was not considered a suitable tool for the present study.

4.2.5 **Network analysis.** The method of analysis discussed in the following paragraphs relates to work described by Ogborn (1979), Bliss and Ogborn (1979) and Mujib (1980) and not to the use of "PS networks" described by Ashmore et al (1979) and already discussed in Section 2.2.3. The latter refers to the analysis of a certain class of problems rather than to the analysis of protocols. Although students' solution routes have been interpreted by comparing them with such networks (Bapat, 1975; McCabe, 1977) more student-centred PS models have been adopted here. Ogborn writes of the difficulty of analysing students' protocols and similar verbal material.

> "We have all learned to live with the fact that the more realistic our data is the more impossible it becomes to say what is there. If we interview children about how they think, we get a strong impression of understanding what is going on rather better, and an equally strong sense of despair at ever being able to capture it without losing its essential complexity."
>
> (Ogborn 1979, p 359)

He goes on to describe the application of linguistics networks to the analysis of such data. This and subsequent relevant work was published only when the present study was nearly completed. Its evaluation was thus in a sense academic and will be reported quite briefly.

The method is most easily described by reference to an example. Mujib (1980) spent two years developing a network to describe some aspects of university students' attempts to solve problems in physics. A modified version, in an early stage of adaptation to the present study, is presented in Figure 4.2. Two conventions are used: a bracket, " { ", representing a logical AND and a bar, " [", representing a logical OR. The network represents a moderately detailed (though in this case tentative and incomplete) interpretation of protocol data from the point of view of an information processing model of PS. It consists of "a kind of generative grammar" which can generate stylised descriptions of the data using only the terms at the ends of the branches of the network. Each of these must, of course, be explicitly defined but also derives meaning from its place in the network. Thus "match" (Figure 4.2) in addition to an explicit definition in terms of comparing two sets of data, is seen from the network to represent the execution of the process of bringing together information; similarly "enough" represents a part

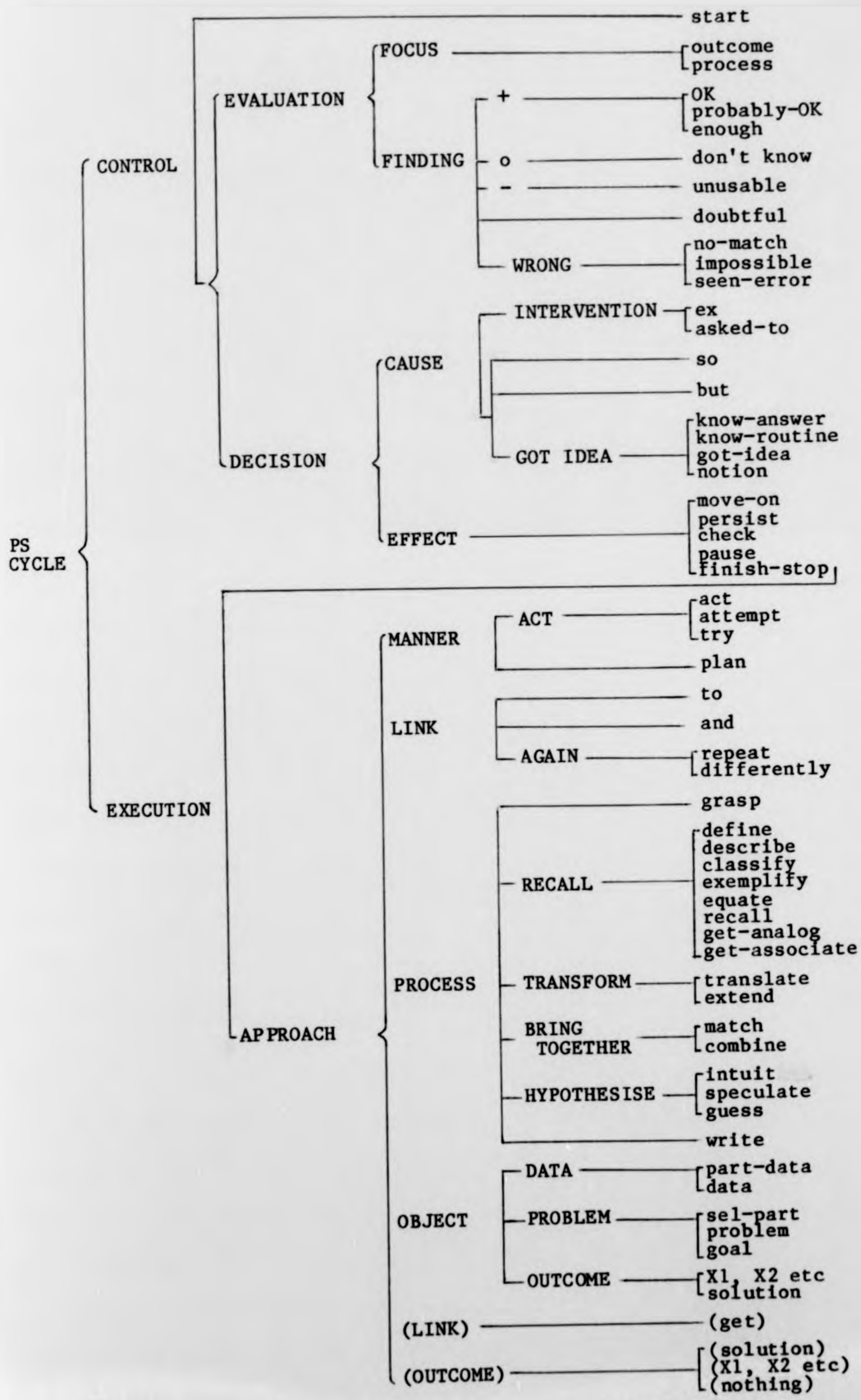


Figure 4.2 Partially adapted network for describing PS protocols (based on Mujib, 1980)

of the PS control system and is the result of evaluating either a procedure or its outcome. Application of the network to the protocol of student 020's solution of item 5a (already discussed in the last section) results in the following coded version:

EVALUATION		DECISION			APPROACH			
			start	act	to	equate	sel-data 1	(get X1)
					and	define	sel-data 2	(get X2)
					and	grasp	problem	
outcome	OK	got-idea	move-on	act	to	define	sel-part	(get X3)
					and	match	data	
outcome	no match	got-idea	move-on	act	to	combine	X3 X1 X2	(get X4)
					and	match	part X4-data	
outcome	enough	but	check	act	to	match	part X4-data	
outcome	enough	so	finish	stop				

This shows that the student first generates new information regarding the two terms used in presenting the tabulated data (number of electrons and mass number) and then (reads and) understands the problem. He next decides to define a term in the problem (isotope) and to match that against the data given. This fails so he logically combines his definition (X3) with all the information generated earlier (X1 and X2) and matches part of the results (X4) with the data given. On evaluation this is accepted as meeting the goal but he decides to check by matching the remainder of the information generated (X4) with the data given and confirms (and writes down) the answer.

If the network presented were to have been used in this study, it would have needed further modification to reflect the model adopted. A great deal of iterative development would also have been needed to ensure that the grammar and syntax represented, defined an adequate language for expressing all the protocols collected. One reason for not pursuing this further was that the analysis was already nearly completed when Ogborn's publication and subsequently Mujib's network came to hand, however it will be of interest to evaluate it briefly against the criteria established for a satisfactory analysis. As regards generality there was no reason to suppose that a single network would not be able to cover all the PS episodes recorded. However, it would obviously be rather an elaborate generalisation, and further iterative analysis would be needed to identify routes through the network in terms of a limited number of strategies. As regards lack of ambiguity, the interpretation represented by a network is arrived at by the same sort of iterative generalisation employed in simple category analysis. However, although no strong claims can be made for objectivity, Bliss and Ogborn (1979) argue that the semantic grammar of the network helps to maintain the stability of the meaning of the categories identified and that it enables sources of disagreement or ambiguity to be pinpointed more easily than in simpler category systems. Finally, as regards its explanatory power, the network already represents a rather detailed interpretation of observed behaviour in terms of a PS model, thus perhaps its greatest strength lies in its explanatory potential. However, Bliss and Ogborn (1979) and Mujib (1980) testify to the complexity and difficulty of the task of constructing a network and Ogborn (1979) to the absence of "criteria for good networks and network features (and) criteria for trading off complexity and subtlety". In the circumstances it was decided, though not without some reluctance, that the more structured and powerful network approach did not offer sufficient advantages in terms of the aims of an exploratory study, to justify a late change from the simple category analysis already adopted.

4.3 The analysis of students' PS strategies and errors

After establishing what is meant by strategy, the section which follows reports the conduct of the analysis, the recording of the data and some general impressions gained during this phase of the work.

4.3.1 A definition of strategy. The word "strategy" is generally associated with military manouevres or games and there is no widely agreed definition in the literature of PS. Butcher (1968) regards a strategy as a plan of action; less general than an attitude or set but less specific than testing a particular hypothesis. This agrees closely with the usage of Bruner, Goodnow and Austin (1956) in their well known work on problems involving concept formation. Ausubel and Robinson (1971) refer to a set of verbal rules which guide the selection and manipulation of information. In the information processing paradigm of Simon and Newell (1972) this would relate to the selection of a particular problem space, that is with the application of a particular, limited set of operators. A pragmatic position was adopted in the present study and strategies were regarded simply as recognisably different ways of processing information which represented students' general approaches to tackling examination items. In terms of the PS stages, strategy is represented by the type of approach selected after a student has read and understood the item (or after abandoning a previous attempt). In terms of memory, a strategy is reflected in the application of a particular set of general procedures in WM.

4.3.2 The conduct of the analysis. A simple category analysis was employed to generate descriptions of strategies and errors in the protocols by an iterative process of analysis and generalisation. Students' protocols were analysed task by task and item by item across the population involved. Initially the broad categories identified during the pilot trials were employed and these were refined, added to and occasionally amalgamated in the light of the data from the protocols. In addition to the large quantity of information to be handled (the typed protocols covered 296 pages), the principal difficulties encountered related to inferring the processes used by students where these were not explicit and to defining boundaries between different categories. Thus, although students generally verbalised fluently after the first few minutes and the protocols provided quite an adequate record of the path of a solution, the processes by which they progressed from point to point were often only implicit. As regards both the inference of process and the definition of boundaries, the PS model adopted proved a useful tool. Even so, many distinctions established early in the analysis were progressively eroded as a range of intermediate cases came to light and the classification which follows cannot be regarded as definitive. However, after several iterations all attempts had been assigned and further analysis would have resulted in only minor changes. It was therefore decided to take stock of the information generated at this point.

4.3.3 **The recording of the data.** The analysis covered 857 separate attempts at solving items (students who were dissatisfied with their first attempt frequently tried again) and included 372 recorded errors and 102 instances of checking answers. In order to facilitate subsequent treatment of the data, a form was developed on which to record all responses to each particular item. A copy will be found in Appendix A. Like the strategy and error categories, the form was developed during the course of the analysis to meet the apparent demands of the situation. Strategies were listed under the following headings:

(i) **Abortive attempts** were recorded only where a student's initial strategy was completely abandoned before any sort of answer was produced. Such attempts were clearly distinguished from principal strategies (see below) which merely got off to a hesitant start, and from cases of failure to make any significant attempt. Only 8% of the strategies recorded fell into this category. By definition an abortive attempt was associated with one or more errors.

(ii) **Principal strategies** represented a student's main attempt at an item and were distinguished from abortive attempts insofar as they led to some sort of answer. There were a few instances in which more than one principal strategy was recorded. These occurred where a student realised that his answer was impossible and returned to make a completely new attempt. Such cases were rare and were distinguished from attempts to simply check an answer in which the student might have more or less confidence. Three columns were provided for recording principal strategies depending upon whether they were (a) correct and sound, (b) correct but unsound and (c) incorrect. By definition the two latter categories involved errors.

(iii) **Checking strategies** were those which followed a principal strategy and attempted to confirm the correctness or otherwise of the answer. They were further sub-divided as cases in which further processing (a) confirmed a correct answer, (b) wrongly appeared to confirm an incorrect answer, (c) wrongly infirmed a correct answer, (d) correctly infirmed an incorrect answer, (o) failed to provide any information concerning the answer and (g) checked, not the correctness or otherwise of the answer, but that it did indeed meet the goal as perceived by the student.

On the form for any particular item, one line was used to record the response of each pupil attempting it. Strategies were indicated by symbols (S1, S2 .....) which will be elaborated in Section 4.4 below. Errors were shown similarly (E1, E2 .....) immediately below the strategy with which they were associated. Separate symbols were used for checking strategies which were analysed separately.

4.3.4 Some general impressions regarding the analysis. In a review of research into the thought processes employed in the solution of verbal arithmetic problems, Hollander (1979) comments that "objective evidence of subjective thought processes" is hard to come by. If we rationalise students' behaviour on the basis of our own interpretive frameworks, is what we achieve an understanding of students' cognitive processes? Ogborn has referred to the "impression of understanding" that we often get from protocol data. But even if our system of analysis were better than it is, that impression may tend to become obscured in a plethora of detail. It therefore seemed important to record a number of general impressions before proceeding.

While no systematic attempt to validate the five stages of PS was either envisaged or attempted, a strong impression emerged that it represented an excellent, broad framework for the description of PS episodes. Possibly linked to this was a very strong underlying impression of "wholeness"; that while pinning down the patterns was a daunting task, there were indeed marked regularities of behaviour underlying every PS attempt. Equally clearly, of course, the part of the pattern illuminated by each attempt depended upon the interaction of the particular test item with the cognitive appartus of the particular student. The impression was thus one of a complex pattern underlying PS behaviour, fragments of which were being illuminated from a range of differing, and not fully known, perspectives. It is, unfortunately, much easier to see that there is such a pattern than to describe it adequately!

Another broad impression was of the very major role played by recall in answering examination items. Both before and after the test, as well as during it, students referred to "revision" (frequently to explain that they felt they were unable to do themselves justice although performance on the test was in fact up to expectation) and it was clear that they expected either to know the answer to each question, or to know a standard routine from which the answer would quickly emerge. Where neither of these conditions obtained, or even where an unfamiliar phrase was employed, the initial reaction in most cases was to want to omit the item or "guess". Considerable encouragement was often necessary to persuade them to make a serious attempt, and guessing or undisciplined speculation often seemed to be preferred to attempting logical or systematic thinking, though this would quickly have resolved the difficulty in most cases. The standard JMB examination rubric, included on the front page of the PS test, states that:

> "A few questions may relate to topics or substances with which you are not expected to be familiar. In such questions the necessary data have been provided and you are being asked to think clearly in a new situation."

However, students seemed disinclined and ill-prepared to tackle anything unfamiliar.

Finally, before passing on to the more analytical part of the report, it should be observed that the sort of analysis undertaken is bound, in some respects, to destroy underlying patterns as well as exposing aspects of them. Although it was necessary to do so to avoid confusion there were, for example, disadvantages in considering strategies, errors and checking separately. In particular the nature of the errors recorded often helps to throw light on the strategies employed. Similarly, if perhaps to a lesser extent, there are losses in considering checking in isolation. Such classifications were used only in the hope that an adequate analysis might facilitate a subsequent synthesis and thus lead to a better understanding of students' PS processes.

4.4 An account of the PS strategies identified

In this section, the nine strategies identified are described. Table 4.4 summarises the names and symbols employed and indicates the incidence of each.

Table 4.4 A list of the strategies identified

Strategy	Symbol	No of cases (total 857)
Omitting	SO	88
Recall	S1	318
Linking generalisation	S1/2	27
Propositional reasoning	S2	267
Analogy	S2'	35
Hypothesis and evaluation	S3	42
Intuition	Si	11
Serial processing	Sp	45
Guessing	SO'	24

In a category analysis of the sort which led to the identification of these strategies, each category is operationally defined by the examples allocated to it. However, it will be more helpful to provide a general description of each strategy and to illustrate this by reference to examples from the protocols.

4.4.1 **Strategy SO - omitting.** SO refers to the omitting of an item and requires little definition. It was used only where a student had read an item with the intention of attempting it but failed to write any answer. Examples are hardly necessary but this description may be amplified a little by noting that, out of 88 cases recorded, 61 followed positive attempts to reach a solution, 20 followed failure either to understand the item or to select a strategy, and 7 followed errors in an earlier related item which rendered any attempt impossible. As regards the relationship of this strategy to the PS model, the interest lies in the reason for a student omitting to answer an item. This will be considered later in terms of errors of understanding, selecting and processing.

4.4.2 **Strategy S1 - recall.** S1 refers to solutions retrieved directly, or with only minor propositional manipulation (as for example in applying a mnemonic) from LTM store. The recalled information may vary from single names to explanations of phenomena involving two or three sentences. It was the most frequently encountered strategy and a number of representative examples are given on the next page. Because recall can be a complex process and because it also plays an important part in all the PS episodes recorded, S1 was sometimes difficult to distinguish from other strategies. Two characteristics of recall which have ramifications for all PS of the type under investigation and which were at the heart of many problems of classification, are its selective and reconstructive nature (Norman, 1973; Lindsay and Norman, 1977). Although the examples in Table 4.5 are deliberately straightforward ones, they provide some illustration of these aspects. In the first example, student 019 knows at once that he knows the answer but there is a short pause while he locates the information in LTM. The search and select aspect of this sort of information retrieval is well illustrated in student 017's attempt on a similar item (Example B). He thinks he knows the answer and recalls studying the topic - then he appears to locate the appropriate general area in LTM and proceeds more confidently - he recalls when the topic was studied - then another experiment done at that time - and finally locates the answer. Examples C to E also illustrate the reconstructive aspect of the recall of more detailed

Table 4.5 Examples of strategy S1 - recall

A. **Student 019; item** 7(c)(i) asked for the name of an industrial process given a technical description -

> "Ah I can remember this it's the er - contact process ..."

B. **Student 017; item** 7(c)(ii) was similar to the previous one -

> "Oh er - name the industrial process er - I remember doing it - nitrogen is er - [speaks more positively] oh third year - fountain experiment - Haber process"

C. **Student 025; item** 3(a)(ii) asked for a chemical test for chlorine -

> "Chemical test - er - well it's a bleacher - so it'll bleach litmus paper - yeah litmus paper"

D. **Student 019; item** 4(c) asked for an explanation of an experimental observation. The student gives a rather rambling account which sounds as if a familiar explanation is being reconstructed; when asked if it was familiar -

> "Yes ... we've done the reaction and we were told the explanation"

E. **Student 012; item** 1(a) asked for the name of each electrode product when dilute sulphuric acid is electrolysed -

> "Well I know it's as it is in water - and - H plus - O minus - so - anode's positive, cathode's negative - so hydrogen at the cathode and - and oxygen at the anode"

Questioning suggested that student 012 recalled at once that the products were hydrogen and oxygen.

information. Student 025 (Example C) recalls the key feature of a standard chemical test first, then adds detail. Student 019 (Example D) knows the explanation demanded but takes a while to put together the different elements. Finally, student 012 (Example E) knows the electrolysis products but has to use additional recalled information to sort out (reconstruct) which gas is discharged at which electrode. This example also illustrates a typical problem in classifying responses. Although the student basically remembers the answer, she has to use a little elementary reasoning to reconstruct one aspect of it. However, as the processing was only of a similar order to that involved in using a mnemonic, the response as a whole was classified as essentially recall; it was of quite a different character from the solutions of a few students who worked out an answer from basic principles.

Because of the central importance of recall it may be useful to analyse this strategy in more detail with reference to the adopted PS model. When the student understands an item it is assumed that he makes it meaningful by translating and encoding it in such a way that he can relate it to his LTM store. There is certainly evidence of such translation in some of the protocols. For example, one of the items already mentioned (Example A, Table 4.5) referred to the oxidation of sulphur(IV) oxide to sulphur(VI) oxide and it was clear that several of the students were unable to continue until they had explicitly translated these names to sulphur dioxide and trioxide, respectively. It further seems reasonable to suppose that the process of understanding an item activates relevant areas of LTM as proposed for example in Atkin's (1978) model. This appears to offer an explanation for the fact that in many S1 solutions students were confident that they knew the answer before actually finding it. Thus, although there was obviously little evidence in cases where an answer was recalled more or less instantaneously, many protocols lent support to the following description of an S1 type solution. The sequence appears to be that the student (i) reads and (ii) understands, translates and encodes the item, activating parts of LTM and realising that the answer should be available "off the shelf"; he then (iii) selects strategy S1 and, (iv) activates it, scanning the relevant area of LTM store and (v) checking what is there against his goal until he finds (or in more complex cases reconstructs) an answer that satisfies him.

4.4.3 **Strategy S1/2 - linking generalisation.** S1/2 refers to attempts which can be regarded as intermediate between direct recall and propositional reasoning. What is recalled is not the answer but a simple generalisation which applies to the case and leads immediately to the answer. Only a limited number of cases was identified and a single example from the protocols will suffice:

F. **Student 017; item** 6(c) referred to a verbally described reaction and asked which of the two reactants was acting as an oxidising agent -

> "... it's got to be one of those surely - and - all metals are reducing agents - so it's got to be copper(II) sulphate"

Examples classified as S1/2 ranged from clear, simple generalisations like the above and "ammonium salt plus base yields ammonia", in which there is formal identity with low-level rule application, to less well defined associations such as "ammonium salts often give ammonia". The essential element is that some aspect of the information given is encoded and classified as belonging to a group about which a relevant generalisation is known. Some cases appeared to be triggered by the information given, others by the goal and some were indeterminate. In the terms of the PS model, the strategy could be described as seeking a generalisation to link givens and goals. The dividing line between recalling a particular case and recalling a generalisation of which the particular case is an example, though formally clear enough, was not necessarily easy to discern from the protocols.

4.4.4 **Strategy S2 - propositional reasoning.** S2 was particularly associated with some of the tasks that had been labelled classification, giving formulae and describing occurrence ("What would happen if ...?" questions). A solution is sought through the recall and application of propositional knowledge, to generate specific information previously unknown to the student (although it may occasionally be recognised later as known but forgotten!). Two relatively homogeneous sets of behaviours covered a majority of examples, namely (i) the generation of selected attributes of a given concept, followed by the identification of particular cases on the basis of processing given or recalled information about them, and (ii) the identification of an appropriate concept, principle or model and the application of propositions stemming from this to given data so as to generate conclusions. Five examples from the protocols which were classified as S2 are given in Table 4.6.

Examples G and H are fairly typical of the first type, involving the application of propositions about oxidation and isotopes respectively. In Example I, the student explains his answer in propositional terms but may have applied a model (involving perhaps electron flow and electron transfer), and Example J represents a more extended version of similar propositional processing which may have algorithmic overtones.

Finally, Example K is unusual and reflects some of the problems of categorisation. Because of the nature of this item the answer was generated in the course of understanding the ideas in the question. The experiment described is interpreted as it is read, applying propositional information from LTM. It was not clear whether the

Table 4.6 Examples of strategy S2 - propositional reasoning

G. **Student 027; item 7(b)(i)** asked for the identification, in an ionic equation, of the substance being oxidised -

> "Oxidation is - loss of electrons - so the Fe - so the H - the hydrogen is positive ... so that means it has - lost an electron - so the (hydrogen) is being oxidised"

H. **Student 020; item 5(a)** asked for the identification of isotopes from a table of data relating to numbers of electrons and mass numbers. After some propositional pre-processing of the data the student continues -

> "Isotopes have the same number of protons but different numbers of neutrons - so they'll have the same number of electrons - but - different mass - number ... look for the same number of electrons - M and Q"

I. **Student 021; item 1(c)(i)** asked about the effect on an electrolysis of increasing the concentration of electrolyte while maintaining constant current -

> "[pause] - well there will be no increased rate of neutralisation [discharge at the electrodes] ... no, because the rate of neutralisation is directly proportional to the current passed"

J. **Student 014; item 5(d)** asked for derivation of the formula of a compound from data about atoms. The student selects data about electrons without comment -

> "I'll have to find the valency of L - 2, 8, 6 - so it's going to be L valency two ... [writes L^{2-}; goes on to deduce that X will be] ... positive with a valency of 1 ... that'll be - 2 negative so'll be - LX_2"

K. **Student 019; item 6(b)** followed a description of an experiment and asked for the name of a product removed by filtration. This was usually generated during propositional pre-processing of the stem -

> "Looks like a displacement reaction OK [continues reading] ... oh the reddish brown precipitate would have been copper ... because the copper was displaced by the iron - the iron's higher up in the electrochemical series"

identification of the precipitate as copper (subsequently recognised as the answer to part (b)) sprang immediately from recognition of the reaction as a displacement type (only marginally on the S2 side of S1/2) or whether the idea of the electro-chemical series actually played a part (positively S2), or represented a check (to be reported elsewhere) or merely a rationalisation for the benefit of the author. Constant interruptions to check up on such points were obviously impracticable; however as the answer clearly involved propositional processing, even though initially selected as a strategy for understanding rather than answering the question, it was classified as S2.

Returning to more typical cases and relating them to the PS model, it seems that students reading a new item may first decide whether it ought to be answered by knowing or by applying knowledge. In the latter case the appropriate propositions are sought in the activated areas of LTM (in Example G these would include not only the definition of oxidation but ideas relating to the charges on electrons and ions etc) and are combined with one another and with given information in WM (Example G " - so ... so ... so ... etc") until a product matching the goal state is reached (Example H "look for the same number of electrons").

4.4.5 **Strategy S2' - analogy.** S2' relates to a relatively small group of attempts in which a pupil used what he judged to be a similar case as an analogy and generalised from the familiar case to the new situation represented by the item. Such a strategy frequently appeared to be "triggered" when such an analogy was perceived, rather than being planned from the start. Examples ranged from the sort of useful, insightful analogy implied in the description, to naive generalisations based on inadequate or false analogies which were sometimes quite vague. At the latter end of the scale they were sometimes little better than guesses. More light will be thrown on cases of this sort when errors are considered. Three examples follow:

L. **Student 018; item** 2(a) asked for the formula of ammonium carbonate -

> "Ammonium carbonate - that's - hydrogen carbonate - H_2CO_3 - there's er - hydrogen's got a valency of 1 - ammonium's got a valency of 1 so that'd be ... NH_4 twice - CO_3 "

M. **Student 012; item** 7(b)(i) asked for the identification, in an ionic equation,of the substances being oxidised. The student struggles with this and then omits it. Later, while attempting another item -

> "it says here **oxidation** of sulphur to - sulphur(IV) to sulphur(VI) so I can apply that ... that's got four elect - four to six - and that's three to two - so I'll put hydrogen is being oxidised"

N. **Student 009; item** 3(d) asked which of three gases normally acts as an oxidising agent -

> "Well the only oxidising agent I can think of at the moment is sulphuric acid - so er - ... - it could be hydrogen sulphide - because that is identical because it also has two hydrogens - like sulphuric acid"[!]

Only a few examples of this strategy used good analogies but on the whole S2' responses were easy to classify. An exception related to the answers to item 6(a) given by all the 18 students who attempted it. The item asked why an excess of reagent was used in an unfamiliar salt preparation which was described in the question. The actual combination of reaction and procedure was obviously new to them and it was felt that students were trying to generalise from more familiar examples.

Although a few cases of this strategy led to processing of the S2 type, the majority mirrored S1 solutions. However, the use of generalisation from analogy meant that all attempts had at least the formal character of propositional reasoning, hence the designation S2'. In terms of the PS model it appears that the student, understanding the item but not recognising it as familiar and unable to proceed in other ways, may recognise an apparently similar case in LTM and attempt to generalise from that.

4.4.6 Strategy S3 - hypothesis and evaluation. S3 relates to attempts at solution involving the generation and evaluation of hypotheses. The strategy was not common and most examples related to one out of only a few items. Its use was obviously restricted to situations in which the student did not know the answer or any routine way to generate it, and thus accords with definitions of productive PS. A good solution appeared to depend upon disciplined or insightful speculation, evaluated in terms of relevant criteria generated in response to the demands of the item. Alternatively, speculation might relate to the goal, seeking to meet criteria related to specific givens. However, the rather limited number of successful answers classified as S3 tended to be represented by protocols in which the first serious hypothesis to be verbalised was the right one, and in which the evaluation was largely implicit. Indeed, few students considered more than one hypothesis and many attempts were characterised by rather vague speculation and poor evaluative criteria. The category was identified by consideration of a range of cases which were otherwise hard to classify, rather than from individual protocols. Examples will therefore be related to items rather than to particular attempts.

O. **Item 8(c)** asked for a chemical test to distinguish between sodium hydrogencarbonate and hydrated sodium carbonate. No student knew an answer; some generated corresponding properties of each to see if they differed; others recalled various testing procedures and tried to deduce the results of applying them to each. Student 008's protocol is a typical example -

> "I'm wondering how you do this - sodium hydrogencarbonate decomposes to sodium carbonate, carbon dioxide and water - hydrated sodium carbonate - is washing soda ... erm - add one to hard water I suppose - which will soften it - hydrated sodium carbonate softens hard water and sodium hydrogencarbonate doesn't"

P. **Item 4(b)** asked for the reason for using a **plastic** beaker in a heat of neutralisation experiment. Some students answered by recall (S1) and one by analogy with a physics experiment (S2'). Others speculated either about the differing properties of glass and plastic which might be related to the needs of the experiment or vice versa. Student 013 is typical -

"Well you're making - producing heat ... heat is being transmitted by the glass - so I should guess the plastic's used to stop this ..."

In terms of the PS model the student, in understanding the item, realises that he does not know the answer nor can he see a way to generate an answer by ordinary propositional reasoning using the information available to him. He may seek but fail to find an analogy. In many instances he may quickly recognise that the only way to answer the question is to try to generate a sensible guess. In a good attempt, disciplined hypotheses would be generated by directed chains of propositional reasoning, the products of which would be evaluated in terms of the demands of the item. However, students seemed reluctant to do this and when forced to it, generally made only one or two quick guesses and only limited attempts at evaluation. More than one volunteered that he would omit such a question for the time being and return to it at the end of the test if there was time. Given the constraints of the examination situation,it would be hard to criticise such a decision!

4.4.7 **Strategy Si - intuition.** Si refers to a student's sudden, unexplained apprehension of an answer to an item. It was infrequently encountered and in every case was associated with items in which S2 was the most common strategy. Si generally represented an early response to an item and was often quite confident; in this respect it can be contrasted to simple guesses (described below) which appear to be made quite fatalistically after some other attempt has been tried. Two examples will suffice:

Q. **Student 008; item** 6(c) asked which of two reagents, in a verbally described reaction, was behaving an an oxidising agent. An S2 solution is considered first but the student cannot see how to proceed and after a few seconds decides suddenly -

"so the copper sulphate **must** be the oxidising agent - because the iron is on its own ..."

R. **Student 026; item 1(c)(ii)** asked about the effect on a simple electrolysis of raising the temperature while maintaining a constant current. The student's protocol starts -

"Ah - I don't think that's got any effect"

Although the answer given by student 008 is a non-typical example insofar as it follows initial consideration of a standard solution, it is one of six cases which could be regarded as arising from naive general impressions (eg that oxidation involves the addition of something) though some cases could equally have been rapid sub-conscious S2 solutions. All of these six cases were subsequently checked by normal S2 approaches. Four of the remaining five cases were immediate negatives on item 1(c)(ii). Although it would be possible to regard these as S1 attempts which failed to find a link in LTM between electrolysis and temperature, three at least were more quickly and positively accepted by the students than would have been expected in such a case, despite subsequent unsuccessful attempts at rationalisation (although student 026 did say that electrolysis had more to do with current than temperature). In view of the limited and uncertain data it seems unprofitable to speculate further about these attempts.

4.4.8 **Strategy Sp - serial processing.** The term "serial processing" was proposed during the pilot trials and was associated with certain inadequate responses. Some attempts represented little beyond a paraphrase of part of the information given, often involving familiar associations and stereotypes, but falling well short of meeting the goal represented by such words as "explain" (about half the cases recorded related to explanations of phenomena). More substantial efforts involved considerable casting about for useful associations and a more directed attempt to process the information given in the general direction of the goal. However, the lack of some essential insight generally led to answers which the students themselves regarded as less than adequate. Three examples should suffice.

S. **Student 023; item 8(b)** asked why soap does not lather in hard water -

> "Oh God - hard water - that's to do with solubilities - no it isn't - hard water - yes it is ... because the soap isn't so soluble in the hard water - and that's why you get - a scum ..."

The protocol continues for ten more lines referring to "substances in the water - and in the soap" as having something to do with it and even naming calcium carbonate. However she is unable to put it all together and adds nothing to the answer quoted.

T. **Student 009; item 2(b)(i)** asked how the student would prove that a colourless liquid was water -

> "Well you could use electrolysis except that'd be a long method - ... it won't be acid or alkali or it'll probably give neutral with - universal indicator ... I suppose there's a lot of reactions you could put down ..." etc

U. **Student 012; item 1(b)(ii)** asked for an explanation of an unexpected experimental result. After a long pause -

> "... experimental error ... perhaps it's that - yes - experimental error - ... it doesn't really seem - [laughs] well - yes - I can't think of anything else"

Serial processing often appeared to be like an unsuccessful attempt at reconstructive recall. In terms of the PS model, the student seems to cast around for appropriate information in LTM. Unable to construct or reconstruct a satisfactory answer, he makes what use he can of the information generated or settles for an associated stereotype.

4.4.9 **Strategy SO' - guessing.** SO' represents a conscious attempt on the part of the student to apply the principle that any answer is better than none! Guesses often followed abortive attempts to apply superior strategies and ranged from completely blind ones to partially informed efforts where the field had been narrowed by earlier processing. Although many guesses were influenced in this and other ways by the information given in the item, a distinguishing feature of "guessing" is that the final answer is selected from amongst alternatives without reference to any principle of chemistry (although criteria such as "well I don't think they'll give you the same gas twice" were not uncommon!). Guessing differs from intuition in being a last resort as opposed to an advance expectation. It differs from serial processing in being confined to items demanding only very short answers and in the essentially arbitrary nature of the final decision. Though students rarely had much confidence in the answers generated by serial processing, these represented a more directed attempt and the solution put forward was the best available in the light of the chemical knowledge they were able to invoke. Three examples follow.

V. **Student 014; item 1(a)(i)** asked for the anode product in the electrolysis of dilute sulphuric acid. The student thinks that he remembers that hydrogen and oxygen are evolved but also considers sulphur dioxide before (mis)guessing between hydrogen and oxygen. No chemical principles are considered at any stage -

> "I think it's hydrogen - ... I'm not sure I might have them the other way round - [writes] - right next one"

W. **Student 017; item 5(a)** asked for the identification of isotopes from a table of data. After a short abortive attempt at propositional reasoning, the student continues, making no reference to his (mis)definition of "isotope" -

> "Looking for some connection between the figures - ... number of electrons - of J is half the mass number - same with L - and M ... but it says which atoms - so it obviously means at least two - I'd say L and M"

X. **Student 024; item 5(a)** is described in the previous example. This student feels uncertain of the definition of isotope and, after several lines of protocol, appears to chose one to fit a perceived feature in the data -

> "... I think isotopes is er - elements with a same mass number and according to the table Q and X - have got the same mass number ..."

The last two examples illustrate the students' tendency to base guesses on some feature in the data and in the last case to misguess a definition to make it fit.

With reference to the PS model, guessing can be interpreted in terms of an earlier breakdown of the PS process and the application of general heuristics concerning the answering of examination items following such a breakdown. The latter are of limited interest as far as the present study is concerned and the breakdown of PS will considered in the next section.

4.5 An account of the PS errors identified

A number of preliminary decisions had to be reached regarding the identification of errors. An error does not necessarily lead to an incorrect answer: the error may for example be such that the right answer emerges in spite of it; two errors might cancel each other out; the student might recognise and correct the error or it might be part of an abandoned attempt. The recording of the students' attempts as "abortive", "correct/sound", "correct/unsound", "incorrect" and "checking" was adopted partly to facilitate decisions as to what should and should not be regarded as an error in the context of the study. It was decided that errors should be recorded if they were incorporated in the response seen by the student as an answer or if they remained uncorrected in an abortive attempt. However, mistakes which students corrected themselves in the course of a particular attempt were ignored, and indeed it would often have been difficult to decide what was a mistake and what was merely laboured thinking. Errors corrected following a subsequent checking strategy were, however, unambiguous and were not excluded. The following listing summarises the position taken as regards each class of attempt:

(i) **Abortive attempts** - these usually represented a student's first approach to an item; they were not followed through and were often tentative. Only positive errors which the student appeared to accept despite abandoning the approach were recorded.

(ii) **Correct/sound attempts** - no errors were recorded.

(iii) **Correct/unsound attempts** - only a few such attempts were recorded and most of these involved lucky guessing; errors were recorded in the remainder however, and included a few cases where mistakes or inadequate information led nevertheless to a correct answer. There were also a few cases in which two or more errors cancelled one another out.

(iv) **Incorrect attempts** - these obviously accounted for most of the errors recorded.

(v) **Checking attempts** - these were further sub-divided to facilitate analysis (see Section 4.6); errors were recorded as in the previous cases.

Nineteen classes of error were identifed and are discussed below under four main headings. Table 4.7 summarises the names and symbols employed and indicates the incidence of each.

4.5.1 **Misunderstanding the item.** Some errors appeared to be associated solely with misunderstanding either the goal (Eg) or the ideas (Ed) in the item. It seems very likely that **inadequate** understanding of one or the other is also a factor in many of the other classes of error reported below.

(i) **Error Eg - misunderstood goal.** Eg refers to attempts which appear to be directed towards the wrong goal. Examples include underlining the product of a reaction when asked to underline the substance being oxidised, naming the product when asked for the name of an industrial process, an ion when asked for a gas and so on. While some were probably genuine misreadings or misunderstandings it seems probable that in others the goal was merely lost sight of in the course of seeking the answer. These two would represent quite different sources of error but unfortunately could not readily be distinguished from the protocols.

Table 4.7 A list of errors identified

Error		Symbol	No of cases
Misunderstood goal		Eg	20
Misunderstood data		Ed	10
Recall failure		E0	54
Incorrect recall -	incorrect recall	E1	48
	inverted recall	E-1	8
	wrong link	E1/2	4
	wrong analogy	E1/2'	11
Ill-selected recall -	inappropriate recall	E1'	37
	incomplete recall	E1/0	49
	misevaluated speculation	E1/3	13
Processing errors -	processing error	E2	28
	inversion	E-2	18
	look-up error	E2f	3
	data selection error	E2d	8
	recording error	E2r	10
	arithmetical error	E2a	2
	rule selection error	E2'	25
Incorrect intuition		Ei	3
Carry-forward error		Ecf	20

(ii) **Error Ed - misunderstood data.** Ed refers to a small number of cases where failure to understand information in the item led to immediate failure in the PS task. Examples included three cases where students (all from the same school) were unable to understand the difference between an actual and a calculated yield, and another where propene, though read correctly, was subsequently understood as propane. The remaining examples involved what might be called erroneous translation of the item as with the student who, asked to predict the effect on an electrolysis of increasing the temperature, based his answer on the assumption that the electrolyte had been concentrated by evaporation making no reference to the direct effect of temperature. All these cases appeared to involve inadequate encoding of the data given, and might be related to corresponding deficiencies in LTM store (although the propane/propene example might have been simply a failure of attention).

4.5.2 **Problems of recall.** Before discussing the eight classes of error considered under this heading, it may be useful to elaborate again on what is involved when a student attempts to recall an answer, or the information with which to construct one. As he reads an item, he translates and encodes the information so that it can be related to semantic (long-term) memory (LTM) and thus understood. Those parts of LTM relating to the subject matter, and to previous experience in answering items, are activated and the item is probably recognised as requiring a particular sort of treatment. This will usually include a requirement to seek particular information within LTM. The accessibility of such information will depend (among such other factors as the frequency and recency of its use) upon the match between the specification of the information required and the way in which the information has been stored. Difficulties can be anticipated if critical aspects of the item are poorly understood and thus inadequately characterised so that the information sought is not clearly specified. Equally, if the relevant information in LTM is poorly learnt and inadequately characterised, it may prove hard to locate and be of limited value when activated. Indeed, poorly differentiated information in LTM may be at the heart of a vicious circle insofar as it leads to poor comprehension of the item in the first place. 60% of the errors observed appeared to relate directly to recall and will be discussed against this interpretive framework under three main headings: A. recall failure, B. incorrect recall, and C. ill-selected recall.

A. **Error E0 - recall failure.** E0 refers to 54 cases in which a particular propositional link was sought but not found. Such errors were recognised when a student made clear efforts to recall some association but failed to find one which satisfied him and which he was prepared to use as, or in constructing, his answer. Such a failure might be attributed to the student not having acquired such a link or to failure to retrieve it. Students frequently, on encountering a recall difficulty, cast around and came up with a number of ideas during their search and therein lay a source of difficulty in interpreting protocols. However, where such a search failed to unearth a link acceptable to the student a clear case labelled E0 was recognised. Such cases were strongly associated with S1 attempts (although they were also encountered in the course of other strategies) and led in most cases to the item being omitted, or occasionally to a guess. For example, student 010 successfully translated sulphur(IV) and sulphur(VI) oxide to sulphur di- and trioxide, respectively, and continued to seek the name for the associated industrial process -

> "... but I don't think I've heard of that - um - no ... I'll leave it"

Where casting around in LTM leads to an incorrect association being accepted (see E1 and other errors below) it may be a moot point whether this also represents an E0 error. The apparent "wrong link" may be an attempt to compensate for a recall failure and find a way of responding to the item. If so, then many cases of E0 were not recorded. Some instances of inferior strategies (Sp, S0') may also follow unrecorded recall failure.

B(i). **Error E1 - incorrect recall.** E1 refers to 49 cases where a student appeared to recall an incorrect association. The clearest cases were those where the pupil made a quick and/or a confident affirmation. Thus student 010, moving on after missing the contact process, asserted at once that the combination of nitrogen and hydrogen to form ammonia was the Bosch process. Cases like this may represent mis-learning or poorly discriminated information storage. More hesitant cases may, as suggested above, involve failure to locate the associative link sought, and the rather hopeful acceptance of an alternative. Thus student 017 knew that "soap is sodium something" and after some hesitation settled for sodium carbonate. In general, cases of E1 appear to involve confusion amongst related ideas.

B(ii). **Error E-1 - inverted recall.** E-1 refers to 8 cases in which information was recalled the wrong way around. Thus student 011 tested for water by watching cobalt chloride go from pink to blue and recalled oxidation as a gain of electrons. Here again the information in LTM appears to be confused and to consist of general rather than well discriminated associations. Although some sort of mechanical error could be postulated to account for the inversion of recalled information, no such mechanism is known nor does it appear necessary to propose one; the far more numerous E1 errors cannot readily be explained in such a way and the interpretation given is preferred.

B(iii). **Error E1/2 - wrong link.** E1/2 is associated with strategy S1/2, linking generalisation, and refers to 4 cases in which incorrect "linking generalisations" were employed. These are very close to E1 and E-1 errors and appear to stem from over-generalised or confused associations in LTM. Thus student 023 wrongly identified hydrogen as the gaseous product of a reaction (item 3(a)(i)) because "hydrogen is given off usually - in most of the reactions - in the mineral acids", and student 021 used potassium permanganate solution to identify chlorine (item 3(a)(ii)) because "an oxidising agent will ... decolourise potassium permanganate".

B(iv). **Error E1/2' - wrong analogy.** E1/2' is associated with strategy S2', analogy, and refers to 11 cases which varied from simple naive assumptions to uncritical generalisations from particular cases. An example of what is meant by a naive assumption was provided by student 027. He produced an (unsound) argument to show that "the amount of oxygen given off" during the electrolysis of dilute sulphuric acid would decrease if the concentration of the acid were to be increased while maintaining a constant current (item 1(c)(i)). However, he finally gave the opposite answer on the basis of a general feeling that "er more concentrated electrolyte - I think the gases would increase ... not that I know the reason". The analogy here belongs to the naive assumption type and is implied and general rather than specific. If the concentration goes up, we get more. A good example of the uncritical use of a more explicit analogy has already been given. Student 009 decided that hydrogen sulphide must be an oxidising agent by analogy with sulphuric acid - they both have two hydrogen atoms! Insofar as such errors appear to involve the confusion or over-generalisation of associations in LTM, they seem to be closely related to the three previous classes of error.

C(i). **Error E1' - inappropriate recall.** E1' refers to 37 cases of the recall of correct but inappropriate information. Examples included cases in which the correct answer was explicitly rejected, as well as some which may have followed failure to recall the required information. The recall and application of inappropriate propositional information during S2 attempts also accounted for a number of cases. Three examples will illustrate this type of error.

Student 011, when asked for a definition of oxidation in terms of changes at the atomic level (item 7(a)), explicitly rejected the anticipated answer relating to electron transfer and substituted a response referring to the proportion of the electonegative constituent. This obviously reflected poor characterisation of the goal which was also exemplified by a number of students who, when asked to describe a chemical test to identify a particular substance, simply stated an ill-selected property without reference to its suitability as a distinguishing test.

The application of inappropriate information in the context of an S2 solution may be illustrated by the case of several students who correctly recalled and applied propositions relating to acid/base reactions to work out the gas evolved when manganese(IV) oxide reacts with concentrated hydrochloric acid - two gave steam as the answer while the others rejected this and made a new attempt. The failure here could be attributed to one or more of a number of causes[1]: failure to recognise the reaction as a whole, miscategorisation of manganese(IV) oxide as a base, inadequate characterisation of the goal so that steam was regarded as an acceptable answer, and so on.

While the classes of error discussed earlier (Section B above) could be interpreted in terms of the retrieval of poorly discriminated associations, this,and the two classes which follow, appear to be the result of poor characterisation of information in the item, particularly as regards the goal state. The criteria for selecting information from LTM may therefore be inappropriate or inadequate. However, as with E1 and related cases, E1' errors may also represent the selection of a "second best" answer following failure to recall the required information.

---

1. Teachers assured the researcher that this should have been a familiar reaction to most of the students concerned, as indeed it obviously was to some.

C(ii). Error E1/0 - incomplete recall. E1/0 refers to 49 cases in which correct but incomplete information was recalled. While the range includes a few cases in which the missing information was sought but not located, in general its absence appeared to be unnoticed and was thought to reflect a poorly characterised goal state. The former included, for example, the recall of an incomplete definition of oxidation which lacked the element needed to enable it to be applied to the item in question. The latter is exemplified by correct statements regarding the conduct of a chemical test which failed to make reference to the observable outcome to be looked for. An attempt to divide the cases concerned between definite recall failure (E0) and selection failure (E1') was abandoned when too many appeared indeterminate on the basis of the protocol data.

C(iii). Error 1/3 - mis-evaluated speculation. E1/3 refers to 13 cases associated with strategy S3, hypothesis and evaluation, and relates to the mis-evaluation of an hypothesis either in terms of its adequacy in meeting the goal set by the problem, or in terms relating it to the given specific data to be processed or accounted for. This error type has clear affinities with the other selection errors discussed above but the nature of the student's strategy draws attention to the evaluative aspect. Thus student 009 speculated regarding the nature of a reddish-brown residue filtered off following the reaction of iron and copper sulphate solution (item 6(b)). He failed to recognise it as a displacement reaction (E0) and was worried about what happened to the copper. Eventually noting that water was present, he settled on copper(II) hydroxide. This answer has obviously been evaluated to the extent that it meets the student's apparent sub-goal of finding somewhere for the copper to go and represents a substance which is insoluble, but the evaluation is inadequate as regards the relationship of the answer both to the data given (there are atoms of hydrogen unaccounted for) and to the goal (the compound is the wrong colour). In the absence of any evidence that such matters were considered, the cause of the error seems to lie in the student's failure to generate adequate evaluative criteria. This obviously relates to the inadequate characterisation of goals or givens associated with the two previous classes of error. However, it was felt that only more detailed analysis than was intended in the present iteration would enable S1/3 errors to be more fully interpreted with any confidence.

4.5.3 **Problems of processing.** Before discussing the seven classes of error listed under this heading, it may be useful to review briefly this aspect of PS. If, in the process of reading and understanding an item, a student recognises that he cannot simply look up the answer but must process information to generate one, he must first select and recall appropriate items of propositional knowledge. These must then be combined with information from the item and with one another in his working memory (WM). Information-processing models generally assume that such processing is serial in nature (eg Atkin, 1978); although this may be an over-simplification, a protocol can in any case present only a linear picture. The detailed relationship between WM and short-term memory (STM) is beyond the scope of the present discussion but it will be assumed that the manipulation of information takes place under the control of STM, the capacity of which is strictly limited. Processing which demands that many separate propositions have to be kept in mind simultaneously may fail if STM capacity is exceeded. If, however, related items of information can be processed together as a "chunk" then the effective processing capacity of WM may be significantly raised (eg Johnstone, 1981). Processing errors, then, can be interpreted in terms of STM overload, although because of the difficulty of establishing exactly what is being attended to and what constitutes a chunk in any particular PS episode, it is not possible to state in advance what might or might not constitute such an overload.

(i) **Error E2 - processing error.** E2 is associated with propositional reasoning (strategy S2) and relates to 28 cases in which propositional information was improperly combined. Although errors of this sort are conceptually distinct from recall errors leading to the use of incorrect (E1) or inappropriate (E1') propositions in an S2 solution, it was not always easy to make this distinction from the protocols and there is some uncertainly concerning a small number of assignments.

Examples classified as E2 errors ranged from one case of a non-sequitur to careless failure to keep in mind some essential fact and to abandoning an item in confusion. The non-sequitur occurred when student 008 apparently deduced that chlorine was formed from concentrated hydrochloric acid because the latter was an oxidising agent. Almost half the cases recorded applied to the two items 2(b)(ii) and 9(b) which asked for the construction of the structural formulae of urea and propene respectively, molecular formulae being given. The errors here typically involved failing to keep in mind information previously generated or given, thus answers were produced in which two hydrogen atoms were missing, in which one carbon atom had six bonds while the rest had four and so on. Some of these errors were corrected on checking and were regarded by the student simply as careless mistakes. A more difficult case to decide (student 014 on item 2(a)) involved the omission of the brackets around the ammonium radical after the formula of ammonium carbonate had been correctly worked out; as he was rushing to get on to the next item this was judged to be a failure of attention rather than a lack of knowledge. A number of the attempts which were abandoned in confusion were in fact difficult to classify insofar as both recall problems and processing load appeared to exacerbate one another (student 014 on item 1(b)(i) and student 018 on item 5(d) are examples). Insofar as processing load appeared to be a factor, they were recorded as E2 errors.

Although some of the errors listed as E2 must be regarded as rather tentative classifications, most of those associated with items involving the manipulation of data relating to atomic structure and the construction of structural formulae (items 2(b)(ii), 5(a) - (d) and 9(b)) can clearly be interpreted in terms of STM overload, that is of losing sight of or failing to keep in mind all the necessary data. This was strongly supported by direct evidence of "rehearsal" in a number of protocols where students were evidently struggling to preserve newly generated information, such as an electron configuration, by repeating it several times before continuing.

(ii) **Error E-2 - inversion.** E-2 refers to 18 cases in which a correctly verbalised proposition apparently became inverted in the course of processing. The majority of examples involved questions relating to redox reactions. Thus student 008 having stated (in relation to item 7(b)(i)) that oxidation is the loss of electrons and that Fe^{3+} gains an electron when it is converted to Fe^{2+}, nevertheless concluded that Fe^{3+} had been oxidised to Fe^{2+}. As with E2, "rehearsal" was a feature of a number of the protocols and STM failure seems a probable explanation for these processing errors.

(iii) **Error E2f - look-up error.** E2f refers to three cases of careless looking-up of data. Thus student 025 on item 5(a) correctly decided that isotopes have the same number of electrons in their neutral atoms, but in looking this up in the table of data provided, selected the atoms with the same mass number. These were interpreted as careless errors, that is, errors associated with failing to keep something in mind.

(iv) **Error E2d - data selection error.** E2d refers to eight cases in which correct processing was applied by students to the wrong data or to only part of the data. Although one or two cases were indeterminate and might have been due to incorrect propositions linking the processes concerned to the wrong data, most were obviously the result of careless failures of attention. Thus for example student 013, in seeking to generate electronic configurations (in the course of attempting item 5(d)), worked on data relating to mass numbers instead of to the number of electrons. The distinction between this error type and E2f is a fine one and may be unnecessary for some purposes.

(v) **Error E2r - recording error.** E2r refers to 10 cases in which a correct response appeared to have been arrived at in the student's protocol but in which an essential part had been omitted on his written answer script. In one case at least, this appeared to be due to uncertainty over the answer and in another an answer adequately explained in the protocol was so poorly expressed on paper that it was marked wrong. These two cases could have involved inadequate evaluation of the information generated in terms of the goal required but might equally have been careless errors. Three extraordinary cases involved recording sodium stearate, sodium carbonate and copper sulphate respectively as stearate, sodium and copper only. Finally, some cases involved giving a reason (rationalisation) for the answer demanded in place of the answer itself or giving only a part of the explanation generated during the protocol. Although careless recording was probably partly responsible it may have been that students were preoccupied by the task of verbalising their thoughts and so paid less attention than they would otherwise have done to what they wrote.

(vi) **Error E2a - arithmetical error.** E2a refers to two instances involving arithmetical errors. Although the items were intended to be non-mathematical, some minor numerical manipulation was involved in a few cases. The two instances referred to were both careless errors in the elementary arithmetic involved in generating the electronic configuration of an atom from the total number of electrons present. Such elementary errors seemed best interpreted as failure to attend adequately to the task in hand, STM being over-occupied in keeping track of other aspects of the problem.

(vii) **Error E2' - rule selection error.** E2' refers to 25 cases in which inappropriate propositional reasoning was applied to information given. The great majority of the instances recorded related to two rather unusual and difficult items (1(c)(i) and (ii)) in which students were asked to predict the effect, on the electrolysis of dilute sulphuric acid, of changing the concentration and the temperature at constant current. The errors in this case generally related to the application of a simple kinetic model without reference to the maintenance of a constant current. While it is possible to interpret both this and other examples of E2' as reflecting a failure to keep in mind an essential part of the information given, it might alternatively be regarded as the selection of inappropriate knowledge in the absence of anything recognised as better in LTM and thus be more closely associated with E1' or E1/0 errors (inappropriate or incomplete recall) rather than with errors of processing.

4.5.4 Other miscellaneous errors. Two other types of error of relatively minor importance were also noted and are reported below.

(i) **Error Ei - incorrect intuition.** Ei refers to three cases of incorrect intuition. Since intuition was recorded as a strategy it was necessary to record incorrect intuitions as errors. The protocols provide no basis for speculation as to the cause of these; however it may be noted that in all three cases the students concerned subsequently checked their answers and found that they were wrong. Nevertheless in two instances the original intuitive answer was written down in preference to that obtained by subsequent processing!

(ii) **Error Ecf - carry-forward error.** Ecf refers to 20 cases in which an item had to be omitted or could not be properly solved on account of an error carried forward from a previous part of the same item. All related to question 3 which was obviously rather unsatisfactory in this respect.

4.6 An account of checking strategies

The model of PS adopted for the present study suggests that students might check their solutions. Instances of such checking are recorded in Table 4.8 which will be employed as a basis for discussion. The table also indicates the results of the checks and the subsequent action taken by the students. The key under the table should be consulted for the interpretation of this information.

As regards the types of check encountered, these will be discussed in terms of the stages proposed in the PS model. They were (i) checks that the answer meets the goal set by the problem as perceived by the student, (ii) checks that the answer is consistent with the student's cognitive map of chemistry and (iii) checks regarding the accuracy or correctness of the processing involved by repeating all or part of the solution using the same or an alternative strategy.

(i) **Error Ei - incorrect intuition.** Ei refers to three cases of incorrect intuition. Since intuition was recorded as a strategy it was necessary to record incorrect intuitions as errors. The protocols provide no basis for speculation as to the cause of these; however it may be noted that in all three cases the students concerned subsequently checked their answers and found that they were wrong. Nevertheless in two instances the original intuitive answer was written down in preference to that obtained by subsequent processing!

(ii) **Error Ecf - carry-forward error.** Ecf refers to 20 cases in which an item had to be omitted or could not be properly solved on account of an error carried forward from a previous part of the same item. All related to question 3 which was obviously rather unsatisfactory in this respect.

4.6 An account of checking strategies

The model of PS adopted for the present study suggests that students might check their solutions. Instances of such checking are recorded in Table 4.8 which will be employed as a basis for discussion. The table also indicates the results of the checks and the subsequent action taken by the students. The key under the table should be consulted for the interpretation of this information.

As regards the types of check encountered, these will be discussed in terms of the stages proposed in the PS model. They were (i) checks that the answer meets the goal set by the problem as perceived by the student, (ii) checks that the answer is consistent with the student's cognitive map of chemistry and (iii) checks regarding the accuracy or correctness of the processing involved by repeating all or part of the solution using the same or an alternative strategy.

Table 4.8 A list of the checking strategies identified

Checking strategy	Symbol	Result of checking *										Total
		A	B	C			D			O		
				∅	1	2	∅	1	2	∅	1	
Check goal	G	2					12	1	1			16
Check memory	M	4^{x}	1							4	2	11
New S1 (mnemonic)	NS1m	1										1
New S1/2	NS1/2	1										1
Repeat S2	RS2	9^{xx}	4^{x}	1			1					15
New S2	NS2	22	1						2	1	1	27
New S1 (reciprocal)	NS2r	16	1				2		1			20
New S2 (elimination)	NS2e	5										5
New S2 (new data)	NS2d	4										4
New S2'	NS2'								1			1
New S3	NS3										1	1
	Totals			1	0	0	15	1	5	5	4	
		64	7	1			21			9		102

Key: * A = correct answer confirmed
B = incorrect answer confirmed
C = correct answer infirmed
D = incorrect answer infirmed
O = unsuccessful checking attempt
∅ = further checking attempted
1 = original answer accepted
2 = new answer accepted
x one check listed is of a part-solution only
xx two checks listed are of a part-solution only

4.6.1 **Check G - checking that the goal has been reached.** Checking that the goal has been reached is implicit in any information processing model of PS. Indeed the progress of PS is regarded as being monitored by the checking of successive information states against the goal until a match is achieved. Some of the PS errors discussed may be associated with failure to keep the goal in mind (that is, some cases of Eg where the wrong goal is addressed, and of E2f and E2r which relate to careless mistakes in looking up and writing down an answer) and others may relate to a weakly or improperly characterised goal (E1' and E1/0 where inappropriate or incomplete information is selected, and E1/3 where a speculation is inadequately evaluated). However, even in the former case, it must be assumed that the student has checked his answer internally against some goal, even if only that of completing the current processing step. Perhaps because such checking is ubiquitous in all PS activity, it may in fact be a fairly automatic process of which the student has often only limited conscious awareness. This is assumed to be the reason that explicit checks of this sort were only rarely encountered in the protocols. Of the 16 cases where a check of this sort was verbalised, only two were confirmatory. Of the remainder, 12 led to fresh attempts, one to an immediate apprehension of a correct answer by elimination and another one to the abandonment of the item (leaving, however, the incorrect answer in place).

4.6.2 **Check M - checking the answer against memory.** Checking the answer for consistency with existing relevant knowledge in memory might obviously be useful and has been discussed before. Again specific references were relatively rare in the protocols and while such checks are by no means a logical necessity[1] it seems probable that, as with

---

1. Atkin's (1978) model, however, incorporates Ausubel's (1978) idea that PS involves meaningful learning. If this is accepted, then some level of compatibility with existing cognitive structure must be established if the new learning is to be incorporated without conflict. However, it is probable that few of the tasks in the test employed would be regarded as PS in Ausubel's sense of the term.

checks against the goal, it was a great deal more common than appeared. Cases recorded varied from one extreme where an answer generated by an S2 strategy was suddenly recognised as correct by specific recall, to the other where a pupil merely seemed to demand that the answer was not unreasonable in the light of what he knew. Five of the 11 examples recorded were confirmatory (including one case where a student claimed to recognise a wrong answer as something he remembered) and six were abortive. Of the latter, four led to further attempts to check the answer.

4.6.3 Checks R and N - checking the answer by reprocessing. 74% of the checks recorded involved a specific check on an answer (or part of it) either by repetition of the original processing (R) or by employing a new approach (N). As Table 4.8 shows, the only examples of type R involved repetition of an S2 strategy and were therefore designated RS2, while type N shows a variety of sub-classes. Type RS2 will be considered first.

(i) **Check RS2 - Repeat propositional reasoning.** RS2 refers to 15 cases in which an answer was checked by repeating the S2 strategy by which it was first obtained. Simple repetition carries the danger of merely repeating an original error, and although the intention is naturally to avoid this, it was observed in four out of the five cases in which an incorrect answer was checked in this way. And of two infirmatory cases, one related to a correct answer! While hesitating to generalise from such scant data, one might have a case for advising students to prefer other lines of checking.

(ii) **New propositional reasoning.** The four groups of checks labelled NS2 represent 56 occasions on which an answer was checked by the employment of new propositional reasoning. Twenty cases, labelled

NS2r, all related to one or other of the three redox items and involved checking by a reciprocal route. Thus student 008, having solved item 7(d)(i) by applying a definition of oxidation to the ionic equation to decide what was being oxidised, checked her answer by starting again and this time working out what was being reduced. Unfortunately, identical processing errors (E-2, inversion) rendered both answer and check wrong. However, three students succeeded in correcting wrong answers through reciprocal checks. Five cases, labelled NS2e, were similar and involved the systematic elimination by propositional reasoning of the remainder of a limited set of alternative answers provided in the item. For example, student 019, having identified two arbitrarily labelled elements as belonging to the same periodic group by generating their electronic configurations from the number of electrons given in a table of data (item 5(b)), continued the process with all the remaining elements to ensure that they could be eliminated. This particular case was quite similar to checks classified as NS2d, which refers to four cases in which new data were processed to provide confirmation (in this case **direct** confirmation) of an answer. For example student 016, having identified isotopes from data relating to the numbers of electrons in neutral atoms (item 5(a)), checked his answer by processing the mass numbers to confirm that the atoms selected also contained differing numbers of neutrons. All the examples in the two latter sub-classes resulted in the confirmation of correct answers.

The 27 cases classified simply as NS2 involved the use of an ordinary propositional reasoning strategy to check an answer which had usually been generated by some quite different approach. Eleven of the answers checked had been obtained by recall, six involved "linking generalisations", five were intuitive answers and two were guesses; only three had been obtained by S2 routes, though these were different from the route by which they were checked. In two cases wrong responses were corrected but most only succeeded in confirming already correct answers.

(iii) **Miscellaneous N-type checks - other new strategies.** The four remaining checking strategies were represented by only one example each and will be dealt with summarily. NS1m refers to the use of a simple mnemonic in checking a recalled solution and NS1/2 to the application of a linking generalisation to check an answer. In both these cases a correct answer was confirmed. Another case, labelled NS2', involved the use by student 021 (on item 7(b)(ii)) of a good analogy based on the previous similar redox item, to resolve the conflict generated when an NS2r check failed to confirm his original S2 solution. Finally, NS3 refers to a rather doubtful case in which a student, rightly dissatisfied with an answer generated by serial processing, speculated about alternative solutions but was unable to evaluate them.

4.7 The relationship between strategies, errors, checks and items

When introducing the analysis of the protocols, reference was made to an impression of "wholeness" and of underlying patterns. Having now reported the analysis, the aim of the last two sections of this chapter is to attempt to shed a clearer light on such patterns through a re-synthesis of the data. The present section explores certain relationships involving strategies, errors, checks and items alone and the final section will bring the students back into the equation.

The discussion is based around a number of tables which explore some of the inter-relationships involved. However, tables and figures may tend to give a false impression regarding the nature of the data being considered. It should therefore be re-emphasised that the underlying analysis was exploratory and impressionistic. While the various strategies and errors described reflect a serious attempt to generalise observations and inferences from the protocols, only a certain proportion of individual attempts could be regarded as typical examples of any particular category and there were many intermediate cases. Thus, although useful formal distinctions have been made, it seems probable that these categories represent identifiable points in a continuum of processing behaviour rather than wholly independent entities (for example, serial processing and speculation/evaluation appear to merge into one another and into guessing, and recall is intimately involved in all the strategies described). In these circumstances, and in view of the problems of interpreting the protocols of untrained subjects, some of the assignments made are inevitably uncertain. Thus, while generalisations regarding strategies and errors represented by large numbers of cases can be made, if not quantitatively, then at least with some confidence, comments relating to other categories must remain fairly tentative.

Tables 4.9.1 to 4.9.4 relate, respectively, strategies to outcomes, strategies to items, errors to strategies and checks to initial strategies. With regard to the incidence of errors it should be noted that (i) a few of the students' attempts contained more than one error and (ii) many of the errors presumed to underlie abortive attempts or a student's decision to omit, to guess or to apply serial processing, could not be inferred with any confidence and were not recorded. Thus there is no direct correspondence between the numbers of incorrect responses (Table 4.9.1) and the numbers of errors (Table 4.9.3).

4.7.1 **A review of strategies.** Table 4.9.1 summarises the outcomes of the different strategies employed. It can be seen that 11% of the PS attempts recorded were abortive and failed to generate any answer at all; 46% led soundly to correct answers and 40% to incorrect responses. Only 2% of attempts led to correct answers by unsound means, most of these being accounted for by lucky guesses or vague generalisations, analogies or intuitions.

Table 4.9.1 The incidence of particular outcomes from different strategies

Strategy	Abortive attempt	Correct/ sound	Correct/ unsound	Incorrect	Total
S0	0	0	0	88/100%	88
S1	38/12%	190/60%	1/0%	90/28%	319
S1/2	1/4%	20/74%	3/11%	3/11%	27
S2	34/13%	149/56%	6/2%	80/29%	269
S2'	1/3%	22/63%	3/9%	9/26%	35
S3	10/24%	12/29%	1/2%	10/45%	42
Si	2/18%	5/45%	1/9%	3/27%	11
Sp	8/19%	0	0	34/81%	42
S0'	0	0	6/25%	18/75%	24
Totals	94/11%	398/46%	21/2%	344/40%	857

Key: the first entry in each element of the table gives the incidence, and the second the proportion, of each strategy resulting in a particular outcome.

Table 4.9.2 (page 133) shows the incidence of different strategies on different items and suggests that there are some clear associations between particular items, or groups of items, and particular strategies. As regards the two most numerous strategies, recall (S1) and propositional reasoning (S2), these can be conveniently summarised with reference to the tasks identified as a basis for constructing a balanced test. The two parts of Figure 4.1 show the items, listed by task, in which more than 50% of the responses were S1 and S2 respectively.

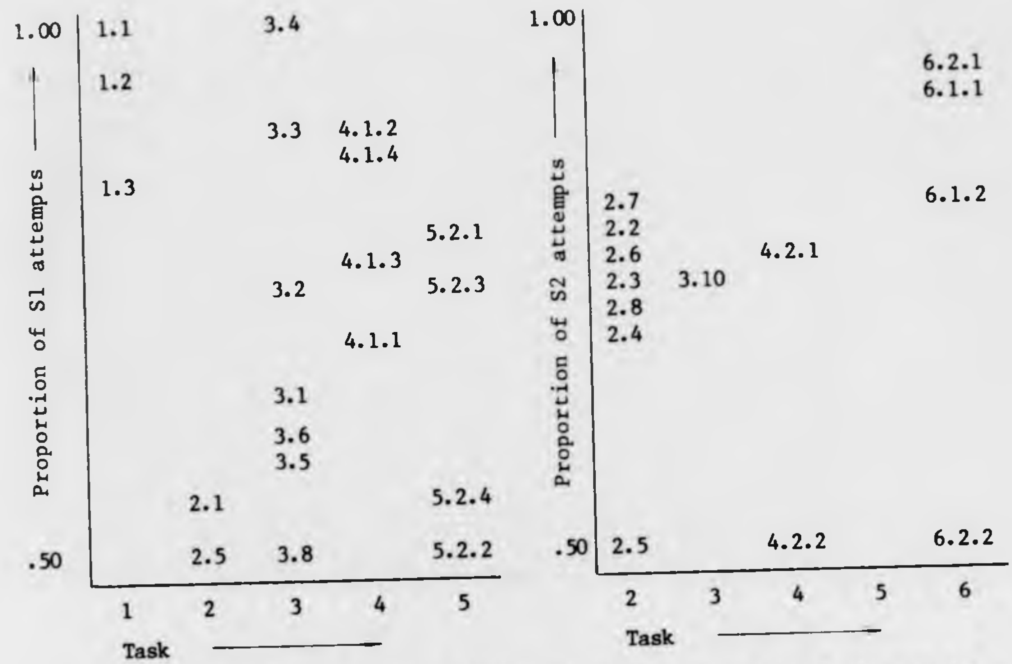


Figure 4.1 Items testing different tasks grouped by dominance of strategies S1 and S2 (scale approximate)

Direct recall was the basis of most responses to items demanding definition (task 1), and to most items involving the naming of a product or process (task 3), describing a procedure or process (task 4) and explaining a phenomenon (task 5); also to two classification items (task 2). Clearly these tasks all related to cases with which the students concerned felt they should be familiar. However, Table 4.9.1 shows that in general only 60% of S1 attempts were successful and the protocols often reflected considerable casting about in an effort to reconstruct an answer. Table 4.9.3 shows that the principal errors associated with unsuccessful attempts at recall were (i) recall failure (E0) in which the information sought could not be found, (ii) incorrect recall (E1 and E-1) which appeared to reflect poorly differentiated information in LTM and (iii) inappropriate (E1') or incomplete (E1/0) recall which may have reflected poor understanding of parts of the item. It should also be noted that, even in items dominated by recall, other strategies were

Table 4.9.2 Incidence of particular strategies on different items

Task	Item	N*	S0	S1	S1/2	S2	S2'	S3	Si	Sp	S0'	Total
1.1	9(d)(ii)	7		7								7
1.2	7(a)	20	1	19								20
1.3	4(a)	11		9							2	11
2.1	3(d)	20	3	12	3	10	1					29
2.2	6(c)	18	2		1	14			1			18
2.3	7(b)(i)	20			3	17	1		1		1	23
2.4	7(b)(ii)	20			3	15	1		1		1	21
2.5	9(d)(ii)	6		3		3						6
2.6	5(a)	20				16			2		3	21
2.7	5(b)	20	3			18					1	22
2.8	5(c)	20	3	2		17					1	23
2.9	9(a)	11		5	2	3	2					12
2.10	9(c)(ii)	7	1	5	1		1			1		9
3.1	7(c)(i)	20	9	20						1		30
3.2	7(c)(ii)	19	5	17						1		23
3.3	8(a)	14	2	14								16
3.4	9(c)(i)	9		9								9
3.5	1(a)(i)	20		12		7					1	20
3.6	1(a)(ii)	20		12		7					1	20
3.7	3(a)(i)	20	3	10	2	7					4	26
3.8	3(b)(i)	20		10		7					3	20
3.9	3(c)(i)	20	2	9	4	5				1	3	24
3.10	6(b)	18		1		16		1				19
4.1.1	3(a)(ii)	20	6	16	1							23
4.1.2	3(b)(ii)	20	3	21								24
4.1.3	3(c)(ii)	20	5	18								23
4.1.4	2(b)(i)	20	1	18						2		21
4.1.5	8(c)	12	2	3				9				14
4.1.6	6(d)(i)	17	7					15		1		23
4.2.1	1(c)(i)	20	1			17	4					22
4.2.2	1(c)(ii)	20	3	1		11	1		5	1		22
4.2.3	4(d)(i)	10	1		2	3	2		1		2	11
4.2.4	4(d)(ii)	10	1		5	2	1				1	10
5.1.1	4(b)	10		3			1	6				10
5.1.2	6(a)	18					18					18
5.2.1	8(b)	14	2	14						1		17
5.2.2	4(c)	10		5		1				4		10
5.2.3	1(d)	20	2	17		1				2		22
5.2.4	1(b)(i)	20	3	14		2				5		24
5.2.5	1(b)(ii)	20	4	10				2		7		23
5.2.6	6(c)	17	4	1			1	9		5		20
6.1.1	2(a)	20				19	1				1	21
6.1.2	5(d)	20	2	1		19					1	23
6.2.1	9(b)	12	1	1		18						20
6.2.2	2(b)(ii)	20	6			13				8		27
	Totals	750	88	319	27	269	35	42	11	42	24	857

* N = number of students attempting any item

adopted by some students. Apart from omitting and guessing, these included serial processing (itself a stepwise effort towards recall) and a few cases of propositional reasoning. Similarly recall was also attempted, with variable success, on many items dominated by other strategies. It was the most frequently adopted strategy and to some extent underlay all attempts insofar as they demanded the employment of recalled information.

Table 4.9.3 Incidence of particular errors associated with different strategies

Error	Strategies									Totals		
	S0	S1	S1/2	S2	S2'	S3	Si	Sp	S0'			
Eg	2	6		12			1	1		20		
Ed		3		4	1					10		
E0		36		9		7		1	1	54		)
E1		22	1	10	2	9		2	3	49	)	)
E-1		5		3						8	)	)
E1/2			4							4	) 72	)
E1/2'				1	10					11	)	) 225
E1'		21		8		4		4		37	)	)
E1/0		36		6	1			6		49	) 99	)
E1/3						13				13	)	)
E2			1	25		1		1		28	)	
E-2			1	17						18	)	
E2f				3						3	) 69	
E2d				7		1				8	)	
E2r		5	1	2		1		1		10	)	
E2a				2						2	)	
E2'				25						25		
Ei							3			3		
Ecf	8	9		3						20		
Totals	10	143	8	137	14	36	4	16	4	372		

It can be seen from Figure 4.1 that propositional reasoning (S2) dominated most of the classification items (task 2), all those demanding chemical formulae (task 6), two "what would happen if ...?" questions (task 4) and one involving the identification of the insoluble product of a particular reaction (item 3.10). Clearly the students employing this strategy recognised that they did not know the answer but that it could be obtained by combining the information given in the item with propositions recalled from memory. Tables 4.9.1 and 4.9.3 show that the success rate for S2 attempts (56%) was quite similar to that for direct recall but that a much wider range of errors was involved. About 40% of these were definite processing errors (E2 to E2a in Table 4.9.3) which appear to be associated with failing to keep in mind some essential element of information and thus probably with STM overload. In addition, a number of errors (Eg) in which the wrong goal was addressed, may speculatively be attributed to students losing sight of the goal in the course of generating the solution. Errors reflected in the selection of inappropriate processing (E2') were quite common but were difficult to interpret. In some cases these may have reflected students' failure to attend to some of the information given, and in others to their inability to make use of it owing to inadequate cognitive maps. Taken together with a fair number of direct recall errors (E0 to E1/0 in Table 4.9.3) it seems probable that, even for attempts based on propositional reasoning, recall errors were as important as processing errors in accounting for PS failures.

The two minor rational strategies involving the employment of linking generalisations (S1/2) and analogy (S2'), respectively, accounted for only a small proportion of PS attempts. However, S1/2 was the most frequently observed strategy on item 4.2.4 (which demanded the elementary application of knowledge concerning the relative properties of weak and strong acids) and S2' on item 5.1.2 (which demanded the reason behind a novel example of a familiar procedure and was not very remote from direct recall). But this accounts for only about half the cases observed and both strategies found occasional application in rather a wide range of items. Both tended to be fairly successful strategies (Table 4.9.1) but were associated with characteristic errors

(Table 4.9.3) designated wrong link (E1/2) and wrong analogy (E1/2') respectively. The latter both appeared to relate to the uncritical generalisation of memorised information to inappropriate cases.

So-called "productive" PS was restricted to a very small number of cases described as S3, hypothesis and evaluation. Nevertheless, it was the most often recorded strategy on three items (4.1.5, 4.1.6 and 5.1.1). These asked respectively for a chemical test to distinguish between two familiar substances, a precaution to be taken in an unfamiliar preparation of a familiar substance and a reason for a particular procedure. The overall success rate for S3 attempts was less than 30% (Table 4.9.1) and many examples so classified were only marginally better than serial processing or guessing. In particular the element of evaluation was often only implicit and appeared to be cursory rather than critical. The errors characteristic of this strategy, designated E1/3 (Table 4.9.3) were associated with inadequate evaluation. Other common errors related to recall failure and incorrect (confused) recall.

The minor strategies labelled "intuition", "serial processing" and "guessing" were fairly well scattered amongst the different items (Table 4.9.2). Insofar as particular errors could be identified, these tended to relate to recall problems. Indeed as regards serial processing, guessing and the relatively large number of cases of "omitting", recall difficulties must be inferred. No attempt has been made to record these owing to the scarcity of information regarding their nature. However, any overall consideration of the present study should keep in mind that well over 100 additional unrecorded cases of recall difficulty should probably be inferred.

4.7.2 An overview of PS breakdown. As regards the breakdown of PS attempts, Table 4.9.3 includes 225 cases (E0 to E1/3) which were identified as recall errors and which can probably be regarded at LTM failures. Similarly, 69 cases (E2 to E2a) were identified as processing difficulties and appear to be associated with STM/WM failure. As regards the remainder, misunderstanding of the given data (Ed) is most closely associated with recall failure, while misunderstanding the goal (Eg) and the wrong selection of processing (E2') probably both contain some examples of LTM failure and some of STM/WM breakdown. There seems to be little basis for speculating either way regarding the few cases of wrong intuitions, and carry-forward errors reflect the nature of the test item rather than of the student's PS behaviour. In addition, it has already been pointed out that there are almost certainly over 100 unrecorded cases of LTM failure associated with the omission of items (S0), with guessing (S0') and with serial processing (Sp). Depending on whether or not these are taken into account it appears that, in the context of tests of the sort employed, LTM failures outnumber STM/WM failures by a ratio of between three and five to one. Interestingly, this is comparable with the finding already referred to, and based on the factor analysis of coded protocol data relating to second year school algebra work, that 50% of the variance on the test employed could be accounted for in terms of the students' "conceptual knowledge" and only 13% in terms of "process components" (Webb, 1979).

Within the area of LTM failures identified, 54 cases were recorded of failure to find the information sought (E0), 72 of retrieval of incorrect information (E1 to E1/2') and 99 of the retrieval of correct but inappropriate information (E1' to E1/3). It is suggested that these are closely related and that, although it is possible that certain critical information may be entirely absent from a particular student's memory store, all these errors can be accounted for in terms of poorly discriminated and organised LTM structure. The student understands an item by reference to LTM, and a poorly discriminated and organised structure may lead to inadequate understanding and thus to inadequate characterisation of the information required. This alone might account for retrieval failure and for the retrieval of correct but inappropriate information. When the information sought is also poorly characterised, then it may be still harder to locate and recognise and, in the absence of clear facts, ill discriminated associations may be interpreted to fit the perceived demands of the problem so as to enable (apparent) progress to be made. Some such explanation must surely underline the surprisingly high incidence of the apparent recall of incorrect facts.

4.7.3 **An overview of checking strategies.** It has already been mentioned that evidence was obtained of all the checking types anticipated in the PS model, namely checks that the answer meets the goals set in the item, checks that the answer is reasonable in terms of how it relates to what the student knows, and checks that there have not been errors in the process of arriving at the answer. Table 4.9.4 shows the incidence of different checking strategies relative to the initial solution strategies which they followed. From this it appears that only 102 checks followed 875 attempted solutions, however this is an over-simplified and misleading picture. Firstly, it has already been noted that checks against the goal state (G) are probably implicit in any attempt at PS, but only explicit cases have been recorded. Secondly, checks against memory (M) as regards the reasonableness of an answer in terms of other relevant knowledge, may also be built into many solutions without being reflected in the protocols. Thirdly, of the 857 attempts, 88 were omits and 94 were abortive attempts with no solution to check.

And finally, certain PS strategies are intrinsically difficult if not impossible to check.

Table 4.9.4 Incidence of different checks following different strategies

Checking strategy	Original strategy						Total
	S1	S1/2	S2	Si	Sp	So'	
G	2		12		2		16
M	1	1	8		1		11
NS1m	1						1
NS1/2	1						1
RS2			15				15
NS2	11	6	3	5		2	27
NS2r			20				20
NS2e			5				5
NS2d			4				4
NS2'			1				1
NS3					1		1
Total	16	7	68	5	4	2	102

If goal checks and memory checks are not considered, on the grounds that too little information is available, then 75 reprocessing checks were made of 675 solutions, which is a proportion of 11%. Even this figure may be misleadingly low insofar as it is unlikely to be practicable for students to check many solutions obtained by recall, analogy, serial processing and guessing. Indeed, only S2 solutions are necessarily checkable and it is noteworthy that the great majority of the checks recorded related to this category. After subtracting abortive attempts, 48 checks were made on 235 S2 solutions, representing a proportion of 20%.

Finally, reverting to the data in Table 4.8 (p 125) and omitting goal checks, the effectiveness of the checking observed will briefly be considered. In 73% of cases a correct answer was confirmed, in 8% an incorrect answer was adjudged correct, in 1% a correct answer was judged to be incorrect, in 8% an incorrect answer was detected for what it was and in 10% the check was not completed. It is interesting to note that rather a high proportion of the answers checked were correct already and that only about half of the incorrect answers were in fact detected as such. Reference back to the protocols shows that in turn about half of these were subsequently corrected. Thus, although further analysis reported in the next section showed that students who performed well on the test were more likely to check their answers than those who did not, checking as observed in this study offered at best a rather modest return.

4.8 A summary regarding students' PS behaviour

One of the aims of the present analysis was to explore the relationship between the PS behaviour of candidates and their examination performance. With this in mind, the section which follows considers the incidence of the strategies and errors associated with different students and the chapter concludes with a general summary of the impressions gained from the first phase of the study.

4.8.1 **Patterns of student behaviour.** Tables 4.10.1 and 4.10.2 show the strategies and errors, respectively, of individual students arranged in rank order according to performance on the test. Because of the small numbers involved, checks have been consolidated as a single column entry in the former table. In addition to recording the raw numbers of cases, Table 4.10.1 shows the proportion of each student's attempt devoted to each strategy, together with a "strategy preference index". This was calculated as the proportion of the students' responses devoted to a particular strategy, divided by the proportion of the population's responses devoted to the same strategy; thus a preference index of 2.50 for student 020 on strategy S2' indicates that this

Table 4.10.1 Incidence of strategies by different students

Student (sex/ school)	Rank	Principal strategies S0	S1	S1/2	S2	S2'	S3	Si	Sp	S0'	Total Principal strategies	Total Checks
021 (M/E)	1	0.00	21(4286) 1.15	3(0612) 1.94	19(3875) 1.24	2(0408) 1.00	4(0816) 1.67	0.00	0.00	0.00	49	10(2041) 1.72
019 (M/E)	2	3(0789) 0.77	13(3421) 0.92	3(0789) 2.50	14(3684) 1.18	1(0263) 0.64	3(0789) 1.61	0.00	0.00	1(0263) 0.94	38	6(1579) 1.33
008 (F/C)	3	0.00	20(4167) 1.12	1(0208) 0.66	19(3958) 1.27	1(0208) 0.51	3(0625) 1.28	2(0417) 3.26	2(0417) 0.85	0.00	48	7(1458) 1.23
026 (M/F)	4=	3(0698) 0.68	14(3256) 0.88	1(0233) 0.74	18(4186) 1.33	2(0465) 1.14	2(0465) 0.95	2(0465) 3.62	0.00	1(0233) 0.83	43	8(1860) 1.56
020 (F/E)	4=	0.00	18(3673) 0.99	2(0408) 1.30	16(3265) 1.05	5(1020) 2.50	5(1020) 2.08	1(0204) 1.59	2(0408) 0.83	0.00	49	12(2449) 2.06
028 (M/F)	6=	2(0465) 0.45	21(4884) 1.32	2(0465) 1.48	12(2791) 0.90	1(0233) 0.57	3(0698) 1.42	0.00	1(0233) 0.48	1(0233) 0.83	43	4(0930) 0.78
027 (M/F)	6=	5(1250) 1.22	15(3750) 1.01	0.00	14(3500) 1.12	2(0500) 1.23	2(0500) 1.02	0.00	1(0250) 0.51	0.00	40	7(1750) 1.47
012 (F/C)	8	3(0638) 0.62	18(3830) 1.03	5(10640 3.38	12(2553) 0.82	3(0638) 1.56	2(0426) 0.87	1(0213) 1.66	2(0426) 0.87	1(0212) 0.76	47	9(1915) 1.61
025 (M/F)	9	4(1053) 1.03	17(4478) 1.21	0.00	11(2895) 0.93	2(0526) 1.29	2(0526) 1.07	0.00	2(0526) 1.07	0.00	38	2(0526) 0.44
024 (M/F)	10	5(1316) 1.28	13(3421) 0.92	0.00	12(3158) 1.01	1(0263) 0.64	0.00	1(0263) 2.05	4(1053) 2.15	2(0536) 1.91	38	2(0526) 0.44
022 (M/E)	11	1(0278) 0.27	12(3333) 0.90	0.00	15(44167) 1.33	3(0833) 2.04	2(0556) 1.13	0.00	3(0833) 1.70	0.00	36	3(0833) 0.70
011 (M/C)	12	1(0233) 0.23	18(4186) 1.13	2(0465) 1.48	13(3023) 0.97	1(0233) 0.57	2(0465) 0.95	0.00	3(0698) 1.42	3(0698) 2.49	43	7(1628) 1.37
017 (M/D)	13	6(1429) 1.39	17(4048) 1.09	1(0238) 0.76	11(2619) 0.78	1(0238) 0.58	2(0476) 0.97	0.00	2(0476) 0.97	2(0476) 1.70	42	2(6476) 0.40
023 (F/E)	14	4(0952) 0.93	16(3810) 1.03	1(0238) 0.76	15(3571) 1.15	1(0238) 0.58	2(0476) 0.97	1(0238) 1.86	2(0476) 0.97	0.00	42	3(0714) 0.60
018 (M/D)	15	7(1892) 1.84	10(2703) 0.73	0.00	14(3784) 1.21	2(0541) 1.33	0.00	0.00	4(1081) 2.21	0.00	37	2(0541) 0.45
016 (M/D)	16	9(2432) 2.37	11(2973) 0.80	0.00	11(2973) 0.95	1(0270) 0.66	2(0541) 1.10	0.00	2(0541) 1.10	1(0270) 0.96	37	8(2162) 1.82
013 (M/C)	17	11(2500) 2.40	11(2500) 0.67	0.00	10(2293) 0/73	2(0455) 1.12	2(0455) 0.93	1(0227) 1.77	3(0682) 1.39	4(0909) 3.25	44	2(0455) 0.38
010 (F/C)	18	8(1905) 1.85	19(4524) 1.22	3(0714) 2.27	9(2143) 0.69	1(0238) 0.58	0.00	0.00	2(0476) 0.97	0.00	42	0.00
014 (M/D)	19=	10(2273) 2.21	15(3409) 0.92	1(0227) 0.72	12(2727) 0.88	1(0227) 0.56	0.00	1(0227) 1.77	2(0455) 0.93	2(0682) 2.44	44	5(1136) 0.95
009 (M/C)	19=	6(1053) 1.03	20(3509) 0.95	2(0351) 1.11	12(2105) 0.68	2(0351) 0.86	4(0702) 1.43	1(0175) 1.37	5(0877) 1.79	5(0877) 3.13	57	3(0526) 0.44
Total (Proportion)		88 (.1027)	319 (.3722)	27 (.0315)	269 (.3139)	35 (.0408)	42 (.0490)	11 (.0128)	42 (.0490)	24 (.0280)	857	102

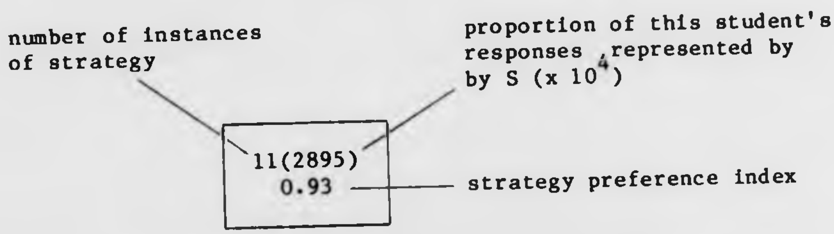
For explanation and key see pages 140 and 143.

Table 4.10.2 Incidence of errors by different students

Student (sex/school)	Errors																			Total
	Eg	Ed	E0	E1	E-1	E1/2	E1/2'	E1'	E1/0	E1/3	E2	E-2	E2f	E2d	E2r	E2a	E2'	Ei	Ecf	
021 (M/C)				1		1		1		1	1	2			1					8
019 (M/E)			1	2					1						1					5
008 (F/C)			2	4				1	2		3	1			1		1			15
026 (M/F)			3	3			1	3	2			1		1			1			15
020 (M/F)		1		1			3	1	1	2	1			1	1			1		13
028 (M/F)	1		2					1	3		2						2			11
027 (M/F)	1		3	1			1		1			1					2		1	11
012 (F/C)		1	2	2				1	4		1	3		1		1	1			17
025 (M/F)	1		3	2			1	1	2	1		1	1				1			14
024 (M/F)			1					2	3			1	1				1		2	11
022 (M/E)	1			1			1	2	4	1	2						2		2	16
011 (M/C)	1		1	2	2			3	3	1	1	1	1		1		1		2	20
017 (M/D)	1	1	2	8				3	2	2	3				1		2			25
023 (F/E)	2		5	3	1	1		1	5	1	3	3					1	1	3	30
018 (M/D)	3	1	7					1	2		2	2		2			2			22
016 (M/D)	3	2	6	2				2	2	2	1	1					2		2	25
013 (M/C)			6	2			2	2	2		3			2		1	3	1	3	27
010 (F/C)	2		1	4	5	1	1	5	4		1				1				2	27
014 (M/D)	3	2	3	3		1		2	1		2	1					2			21
009 (M/C)	1	2	6	8			1	5	4	2	2			1	3		1		3	39
Total	20	10	54	49	8	4	11	37	49	13	28	18	3	8	10	2	25	3	20	372
Proportion	.0538	.0269	.1452	.1317	.0215	.0105	.0296	.0996	.1317	.0349	.0753	.0484	.0081	.0215	.0269	.0054	.0672	.0081	.0538	

student was 2½ times more likely than average to employ this strategy.

A key to the entries in Table 4.10.1 is given below:



Key to entries in Table 4.10.1 (p 141)

Owing to the uncertain reliability of some of the assignments, the students' preference indices were regarded as indicators of trends rather than as quantitative information, and particular caution was observed regarding some of the minor strategies where the indices would obviously be very sensitive to even one or two mis-assignments. For this reason the calculation of similar indices regarding errors was quickly abandoned. With only 372 entries to be distributed amongst 380 cells in Table 4.10.2 (compared to 857 in 180 cells for strategies) most of such indices would have reflected only one or two cases of error and it was felt that the raw data would provide a more realistic picture. For similar reasons, no statistical interpretations of the data are reported.[1]

The students' strategy preference indices can obviously vary upwards from zero with no theoretical upper limit. The two most important strategies S1 and S2, however, show a remarkably narrow range of indices (0.7 to 1.3) suggesting that students differed relatively little as regards their tendency to select these general approaches. For other strategies the indices tend to vary rather more widely, though they are obviously less

1. Some statistical analyses were undertaken with respect to the strategies and checks. While these tended to confirm the impressions reported here it was impossible to arrive at any reliable estimate regarding significance levels.

reliable on account of the smaller numbers involved. However, attempts to arrive at generalisations regarding the data in these two tables met with very limited success. Similarly, an attempt to produce students' strategy profiles recording apparently high and low indices led to no significant conclusions and will not be reported in detail. In particular, no obvious differences in strategy preference profiles were found between students of high and low performance on the test, except insofar as the low-performing students showed a tendency to omit more items, to make more use of serial processing and guessing and to make less use of checking; but even these were relatively weak trends with many individual exceptions. Conversely the main rational strategies (S1, S1/2, S2 and S3) showed a slight trend in favour of students with higher performance. These trends were mirrored as regards errors, with students of lower performance showing a tendency to make more errors in most categories, rather than making different kinds of errors. Thus, as regards recall, the four most prevalent error types (E0, E1, E1' and E1/0), all showed weak trends in the direction indicated. This was also true for main processing error (E2). Weaker students also appeared to be more likely to misunderstand an item (Eg and Ed) and to suffer the effects of carry-forward errors (Ecf).

Two other possible sources of systematic PS differences were also considered; namely school and gender. Although details of individual PS attempts certainly appeared, from the protocols, to reflect different learning experiences in different schools, no general trends were apparent. Some trends relating to gender were noted however, although with only five girls in the sample, and in the absence of statistical tests, these must be regarded as impressionistic only. As regards strategies, girls appeared to be much less inclined to guess than boys (mean preference index 0.15 compared to 1.23) and, although the number of cases was very small, possibly more inclined to use intuition (mean preference index 1.67 compared to 0.78). As regards PS difficulties girls may be somewhat more inclined to processing errors (E2 to E2a in Table 4.10.2) averaging 4.4 instances each on the test as compared to 3.1 for boys. However, it is difficult to assess the significance of these impressions and it is beyond the scope of this phase of the study to speculate further.

Although generalisable trends were markedly weak or absent, some interesting individual differences may be discovered from Tables 4.10.1 and 4.10.2. But even these appear to be merely idiosyncratic and no general strategy and error profiles were identified. Thus it is difficult to know, for example, what significance to attach to the fact that student 010 accounted for five out of the eight recorded cases of inverted recall (E-1) or that student 012 was more than three times more likely than the average student to make use of a linking generalisation. Although these and similar observations seem likely to be attributable to personality and learning experience variables, it was impossible to pursue such matters on the basis of the data available.

4.8.2 **General conclusions.** The various analyses described in this chapter and the subsequent synthesis attempted in exploring the inter-relationships of strategies, errors and students, have enhanced the author's strong impression of the homogeneity of PS at this level. This impression is based on the apparent validity of the five stages of PS as a useful descriptive framework and on the major role of recall and routine. It appears that "translation" of an item with reference to students' own frames of reference, followed either by (i) selective, and often reconstructive, recall or by (ii) the application of a recalled or reconstructed routine, represents the normal basis of answering examination questions. It may be speculated that strategies other than those labelled S1 and S2 represent a fairly limited range of responses to difficulties encountered during these basic procedures. The emphasis seems to be firmly on an LTM structure interacting with the test item at a relatively superficial level with little non-routine activity beyond a general "casting about" to try to reconstruct information not immediately available. It seems probable that well over 80% of students' PS errors can be directly attributed to LTM defects, processing errors being of relatively limited importance. And while processing errors can be interpreted in terms of WM/STM failure, this in turn may sometimes reflect attempts to handle information retrieved from LTM in a poorly organised state (see, for example, Johnstone (1981) regarding "chunking").

Waterman and Newell (1971) have stated that "protocol analysis exhibits a lack of formality and an inherently inductive character that seems to characterise much other scientific (and real world) activity". In the present study it has provided information regarding students' PS behaviour in an examination context which would have been difficult to obtain in any other way. Although the evidence remains impressionistic, the central role of LTM organisation would probably come as no surprise to most teachers. The dominance of the two major strategies and the very narrow range of the students' strategy preference indices for these (Table 4.10.1) provides more quantitative support for this point of view, as does the lack of correlation of PS performance with general intellectual ability. Consequent decisions regarding the second phase of the study are described in the next chapter.

CHAPTER 5

THE PLANNING AND ADMINISTRATION OF THE SECOND PHASE OF THE STUDY

This chapter is concerned with the planning and administration of the second phase of the investigation. This was undertaken in the light of the insights gained from the exploratory analysis of students' problem-solving behaviour reported in the previous chapter. The alternatives considered are briefly reviewed in the first section. Two areas selected for further investigation are discussed in detail and reference is made to relevant literature. The detailed planning of the second phase of the study is then described and the chapter concludes with an account of the design and administration of the instruments used.

5.1 Decisions regarding the direction taken in the second phase of the study

As already indicated, the analysis of students' responses to the short-answer non-mathematical items administered in the first phase of the study was exploratory and impressionistic. It would have been possible, during the next phase, to attempt a deeper level of analysis and perhaps to broaden the scope of the study to embrace mathematical and objective items. However, without further development of the methodology of protocol analysis, it was uncertain how fruitful such work would prove and it was decided to employ independent methods to extend the study. In particular it was decided:

(i) to examine the hypothesis, suggested by the first phase of the study, that it is the organisation of chemical knowledge in long term memory (LTM) that differentiates the more successful from the less successful examination candidate, and

(ii) to explore the influence on PS behaviour of aspects of task formulation, specifically the quantity of information provided.

5.1.1 **The organisation of chemical knowledge in LTM.** A number of studies has drawn attention to the importance of LTM as mediating between learning and PS performance (Mayer, 1975; Gagné and White, 1978). Johnson (1971) has spoken of the need to investigate the role which knowledge structures play in PS behaviour, whilst Cochaud (1979) has provided evidence of the idiosyncratic nature of such structures in college science students. Kempa and L'Odiaga (1983) have reported that performance on examination items in chemistry is distinctly more influenced by their subject matter content than by the intellectual abilities and skills they test. This is in agreement with Wason and Johnson-Laird's (1972) arguments that "content is crucial" and that "any general theory of human reasoning must include an important semantic component".

The analysis of students' PS protocols described in Chapter 4 suggested that there was relatively little difference between high and low performers as regards either the strategies that they employed or the nature of the errors that they made. As there was also no significant correlation between PS performance and general ability as measured by the AH4 test, it was hypothesised that a student's relative success or failure might be related to differences in the way in which knowledge of specific chemical topics is internalised as an organised structure in LTM. This hypothesis was investigated using a word association test to map aspects of students' cognitive structures relating to examination items which they attempted. The nature of the construct "cognitive structure" is discussed in Section 5.2 below, which also includes a selective review of methods which have been used to map it and the rationale underlying the decision to use word association tests for this purpose.

5.1.2 The influence of task formulation. An analysis of the different ways in which the task represented by a particular examination item may be formulated, is presented in Section 5.3. In particular the quantity of information provided is identified as an important variable. Examination items typically present candidates with idealised PS tasks, whereas in the real world the problem solver may have to contend with conflicting, redundant and missing information. Krutetskii (1976) used items involving both missing and redundant information to investigate students' PS behaviour in mathematics but to date there have been no similar studies in science. The influence of missing and redundant information on PS behaviour was therefore investigated as a first step towards bridging the gap between PS in chemistry examinations and PS in more realistic chemical situations.

5.2 The mapping of cognitive structure

In describing students' PS behaviour, reference was made to a simple model of the structural components of memory (Figure 2.3.2).
This showed the interrelationships of short-term memory, working memory and long-term memory. The term "cognitive structure" will be used henceforth to refer to the contents and organisation of an individual's knowledge-base in long-term memory. This is compatible with the definition of Shavelson (1974), who described cognitive structure as "a hypothetical construct referring to the organisation (relationship) of concepts in memory" and that of Wierstra (1979) who sees it as "an organised network of interrelated concepts which a pupil stores in his memory".

White (1979a) has proposed four interacting aspects which affect descriptions of cognitive structure: "purpose", "model and unit", "methodology" and "dimensions". In the present study, the "purpose" of the descriptions sought is to aid in the interpretation of differences in students' PS behaviour. The "model" employed is a semantic network and the "units" are concept labels and semantic relationships between them. The description of LTM structure already given in Section 2.3.1, refers to propositional, algorithmic and heuristic information. However, only propositional information is fundamental insofar as algorithmic and heuristic procedures can be expressed in propositional terms. Propositional information, in turn, can be expressed in terms of semantic relationships between concepts.[1] "Methods" employed in examining network models of cognitive structure are reviewed below with particular reference to word association testing which has found considerable application in science education and which was adopted for the present study. Indeed, Preece (1976c) has gone so far as to provide an operational definition of cognitive structure in terms of the data obtainable from such tests. Finally "dimensions" have been proposed for providing summary descriptions of particular aspects of cognitive structures, usually in terms of numerical indices. However, these were felt to be less important, for the purposes of the present study, than more detailed qualitative descriptions and will be considered only briefly.

The discussion which follows starts with an account of recent reviews relating to the mapping of cognitive structure. A number of paradigms

---

1. The model adopted does not refer to memory for images and episodes as proposed by Gagné and White (1978). However, students' protocols suggested that such elements were not of great importance in the context of answering examination items in chemistry.

are identified and relevant work in each is then reviewed and discussed. The discussion is extended, in the case of word association testing, to embrace methodological and data analysis issues.

5.2.1 Techniques for investigating cognitive structure. Investigations of cognitive structure in the context of science education have made use of a number of different techniques. In a comparative evaluation of methods Shavelson and Stanton (1975) examined word association, card sorting and tree-graph construction methods. In a similar study, Preece (1976a) referred to the use of similarity ratings in addition to these three. He also drew attention to the widespread use of problem-solving tests to investigate aspects of cognitive structure. In a critique of methods for the investigation of cognitive structures, Stewart (1979) referred to these paradigms and also commented on digraph analysis. While he did so in the context of its application in deriving subject matter structures from textual material, this method has also been employed by a group at the University of East Anglia in assessing students' cognitive structures on the basis of the written materials which they produce (for example Pereira 1976, 1979; Maskill and Pereira 1979). More recent reviews have referred to growing interest in clinical interviews and related methods (White, 1979a; Sutton, 1980; Fisher, 1979). Before proceeding to a detailed review, it will be convenient to consider briefly the broad characteristics of the various techniques under three headings: problem-solving, clinical methods and semantic proximity methods.

(i) **Problem-solving.** Problem-solving tests have been used in science education to yield implicit information concerning cognitive structure (Preece 1976a). It has, however, been noted that the intrusion of information processing into such tasks renders this approach suspect (Deese, 1965; Johnson, 1969). Preece (op cit) has described PS tests as "too drastic as tools for the initial exploration of cognitive structure" and has directed attention to techniques "which tamper less with what they are seeking to measure". For similar reasons attempts to assess cognitive structure through modified examination items (Ring and Novak, 1971) or school tests (White, 1979a) should be interpretd with some caution. In the context of the present study, the use of PS tasks to explore cognitive structure would, in any case, have involved a tautology since the purpose of elucidating the structure was to interpret PS behaviour. This method will not, therefore, be discussed further.

(ii) **Clinical methods.** Individual clinical interviews, often associated with paper-and-pencil or practical tasks after the style of Piaget, have recently come to prominence in connection with interest in elucidating children's existing ideas ("alternative frameworks") about science. Pines et al (1978) have contrasted inflexible and flexible interview formats in this connection, and White (1979a) has distinguished between general interviews covering a substantial body of subject matter and restricted interviews which have a much narrower focus. Other workers have described situations in which direct observation of students performing laboratory tasks plays a major role (Fisher, 1979; Driver, 1981, 1983). Although some of these studies have led to attempts at mapping, they relate to work which was contemporary with, or which post-dates, the present study. To the extent that such methods tend to involve problem-solving and the interpretation of children's verbal protocols, there would also have been an element of tautology in employing such methods. For both these reasons only a brief review is given in Section 5.2.2 below.

(iii) **Semantic proximity methods.** There are five methods which broadly fall into this category: similarity rating, card sorting, tree-graphing, digraphing and word association testing. All methods under this heading employ words representing concepts and try to establish how closely these words are related in a subject's mind. In his comprehensive review of research on the organisation of scientific concepts in semantic memory, Preece (1978b) writes:

> "In mapping cognitive structure, the aim is to obtain structural information about the internal representation of concepts. The concern is with the second-order structure involving the relations among concepts in semantic memory rather than the first-order relations within each internal representation (concept), and for this reason the term "map" seems particularly appropriate. Thus, in a map of the world, the distances between cities are represented, but the internal structures of the cities are not. Nor is the neurophysiological basis of cognitive structure sought; in computer terminology, the aim is to elucidate the format of the data base but not to discover how this is is embodied in the circuitry.
>
> Just as geographical maps are concerned with distances between places, maps of cognitive structure are based upon semantic distances between concepts."
>
> (Preece 1978b, p 548)

Sutton (1980) has noted the limitations of the mapping analogy as applied to some aspects of cognitive structure, however his reservations relate to the representation of "dynamic aspects of mental organisation" rather than to the organisation of information in LTM.

Semantic distance, or its opposite semantic proximity, is operationally defined by the methods used to assess it. It is intended only to outline such methods at this point. In the **similarity rating** method subjects are asked to use a rating scale to indicate the perceived similarity of given pairs of concept words. A variety of **card sorting** methods have been employed for the same purpose although these have been little used in studies involving science. Subjects may be asked to sort cards bearing the names of key concepts into as many piles as they think appropriate on the basis of their judgements of similarities and differences. Alternatively, they may be asked to position gummed labels on a large sheet of paper in a way which appears meaningful to them. A similar procedure is employed in the **tree-graphing** method, but subjects are constrained by instructions which result in the construction of tree-graphs reflecting a hierarchy of perceived similarities. **Digraphing** (directed graphing) is based on the parsing of written materials produced by the subject in response to a variety of different instructions. A series of formal rules is then used to reduce each parsed sentence to a graphical form consisting of concept words linked by arrowed lines representing directional relationships. Finally, the most widely employed method in the field of science education has been the **word association** test. Subjects are presented with a series of words representing key concepts and are asked to write down the first word that comes to mind on reading (or hearing) each key word. A number of variations of the method is possible but most require subjects to make a number of associative responses to each stimulus word.

For each of the five procedures described above, methods have been devised for converting the data collected into tables of numerical indices representing inter-concept semantic distances or semantic proximities.[1] Such tables may then be interpreted in a variety of ways. In particular three main approaches have been used to generate corresponding maps reflecting cognitive structure (Preece 1978b). **Spatial maps** may be generated by a multi-dimensional scaling technique (Kruskal 1964). This produces a solution in which concepts are represented as points in n-dimensional semantic space (where n is an integer determined during the analysis) such that the inter-point

---

1. Data from word association tests also lends itself to useful descriptions relating to word meaning and similar dimensions. Some of these are discussed further in Section 5.2.5.

distances are related to the original inter-concept indices. Two-dimensional **hierarchical maps**, which closely resemble tree-graphs, can be obtained by a number of cluster analysis techniques (see eg Everitt, 1974). As each method incorporates its own mathematical model, some caution is needed both in selecting the method and in interpreting the resulting map. Non-hierarchical clustering methods have also been described but have not been used in science education research. Finally, a **graphical** solution described by Waern (1972) has been employed in recent studies relating to science. This reduces a matrix of semantic proximities to a two-dimensional graphical network.

Studies which have used semantic proximity or distance estimates in the context of science education are reviewed in Sections 5.2.3 (digraphing), 5.2.4 (similarity rating, card sorting and tree-graphing), and 5.2.5 (word association testing) below.

5.2.2 **Clinical interviews and related methods.** A number of studies, particularly in the USA and Australia, have used individual interviews to map cognitive structure (eg Rowell, 1975; Pines, 1977; West, Fensham and Garrard, 1982). Similar work elsewhere, using this general methodology, has tended to concentrate on students' understanding of particular concepts rather than on mapping structure (eg Gilbert and Osborne, 1979; Andersson, 1979; Osborne, 1980; MacGuire, 1982) and will not be considered further.

Pines and co-workers have focused much attention on techniques of interviewing (Pines, 1977; Pines et al, 1978). They distinguish between rigid and flexible interview formats and favour the latter to generate a "rich and useful database". In a "totally flexible format" a student is presented with a task and asked a prepared question.

> "Regardless of what the task itself is or what the subject's responses may be, the interviewer is obliged to follow up every response. The subject is never interrupted while responding and his or her remarks are decisive in the course the interview will take."
>
> (Pines, Novak, Posner and Van Kirk 1978, p 8)

The resulting student protocol is transformed into a set of propositions and thence into diagrammatic form. Concepts are mapped onto points and relationships between them on to a linking network of arrowed lines. Maps of similar appearance can also be produced by methods proposed by Stewart (1980). These include line labelling tasks used in conjunction with card sorting or tree construction methods, and sentence generation tasks. Card sorting methods, associated with individual interviews, have also been used by Champagne, Klopfer, Desena and Squires (1981) to produce "structural representations" of students' knowledge.

An introduction to some of the work using clinical methods in Australia is given by West (1979). More recently White and Gunstone (1980) have described the use of "restricted" interviews (ie restricted to a single topic) to describe students' cognitive structures in terms of scores on nine a priori dimensions labelled "extent", "precision", "internal consistency" and so on. Five quite different dimensions are proposed by West and co-workers in an unpublished paper relating to on-going work (West et al, 1982). They also describe the production of an interesting series of progressively elaborated maps, based on concepts linked by labelled arrows, which can be constructed on the basis of group tests followed by relatively short individual interviews.

Two reports published recently in this country relate to attempts at mapping based on tape recordings of students during the performance of laboratory tasks (Fisher, 1979; Driver, 1983). Driver's observations were made in the course of an ordinary science lesson and her model is descriptive and dynamic, rather than analytical. It represents episodes of thinking rather than the organisaton of concepts. It may reflect the activities of working memory but it does not attempt to describe the structure of LTM. However Fisher's work, based on written as well as spoken responses to specially designed practical tasks, led to the construction of a composite map for a group of students in a form similar to that adopted by Pines et al (1978), Champagne et al (1981) and West et al (1982).

Work in the clinical interview paradigm is still in an early stage as regards its application to cognitive mapping and warnings continue to be sounded concerning problems of interviewing and interpretation (White, 1979b; Pines, 1980b; Posner and Gertzog, 1982). The latter authors have noted, in particular, the danger of being misled by responses which have been labelled "suggested convictions".

> "A reply that is stimulated or suggested by the questioner's choice of words or sequence of questions, or one that is given in an effort to satisfy the examiner, is a suggested conviction."
>
> (Posner and Gertzog 1982, p 197)

More generally it may be objected that there is a danger that, in responding to the demands of the interviewer, the student's active selection and manipulation of ideas may distort the picture which emerges of his underlying cognitive structure.

5.2.3 **Digraph analysis of written materials.** Directed graph (digraph) analysis has been used to represent the structure of science concepts in passages from standard textbooks (Shavelson, 1972; Shavelson and Geeslin, 1975) and in defining relationships (Preece, 1976a, 1976c). As previously noted, the method involves the reduction of each sentence or relationship to a diagram in which key concepts are linked by directed lines. The rules for producing such diagrams will be found for example in Shavelson and Geeslin (1975); modifications have recently been proposed by Pereira (1979). The digraphs may be directly combined to yield a graphical representation showing the interrelationships of concepts in a single diagram (Preece, 1976a, 1976c) and/or may be converted to a distance matrix by counting the shortest routes between all pairs of concepts.

A number of studies at the University of East Anglia have used digraph analysis of students' essay material in an attempt to elucidate their cognitive structures (Johnson, 1976; Dangerfield, 1977; Pereira, 1976, 1979). In addition to some technical criticisms, the problem exists that digraph structures may reflect the objectives of the writer rather than his cognitive structure, and that ideas may be deliberately manipulated and ordered in the process of composition (see discussion appended to Maskill and Pereira, 1979). That students are actively selective about the order in which they present concepts in essay materials, seems to follow from further work reported by Pereira and Maskill (1983). It has also been objected that digraphing reflects the syntactic rather than the semantic features of a communication, and Stewart (1979) has listed this and seven other limitations which he believes call into question the usefulness of this technique.

5.2.4 Similarity rating, card sorting and tree-graphing. These methods represent fairly direct attempts to obtain information about a subject's perception of the semantic proximity of given concepts.

A simple rating scale was used, alongside word association testing, by Johnson (1969) to monitor changes in students' perceptions of the similarity of certain physics (mechanics) concepts, following study from a textbook. The method yielded results similar to those obtained from the word association test and was employed again by the same author in exploring the problem of assessing "perceptual structures" of scientific knowledge (Johnson, 1971). He concluded that such procedures can be used to index "mastery of a domain of knowledge" but that the adequacy of the approach still needed to be tested by further research. Kass (1971) has used students' similarity ratings to investigate structure in perceived relations among 20 physics concepts and Fenker (1975) used similar methods to investigate students' cognitive structures relating to elementary statistics. Cachapuz (1979) found that similarity ratings yielded similar structures to those obtained using word association tests in chemistry. The method has also been used to provide empirical data concerning students' concepts as a basis for revising instructional units in physics (Wierstra, 1979).

In a study directed towards the "validation of interpretations of three measures of cognitive structure", Shavelson and Stanton (1975) have described a simple card-sorting task. Twelve mathematics teachers were asked to sort key mathematical (operational systems) concepts printed on cards into as many piles as they thought appropriate on the basis of perceived similarities and differences. All pairs of cards in the same pile were given a similarity rating of one, other pairs being rated zero. By combining the results, a group similarity matrix was constructed with cell values which could obviously vary between zero and twelve. Hierarchical cluster analysis of the matrix yielded results similar to those obtained by two other measures, although the card sorting methods produced three major clusters rather than the two produced by the other methods. Fisher (1979) refers to the use of a similar procedure but does not provide full details of the method used nor of the results attained.

The other methods used in the Shavelson and Stanton study, were a word association test and a tree-graph construction task. The latter has been fairly widely used in research in science education (Shavelson, 1974; Preece, 1976a; Fisher, 1979; Moynihan, 1982). Subjects are given cards bearing the names of concepts and are asked to select the two which appear to be most closely related, place them together on a large sheet of paper and to join them with a line. A third concept may then be joined to the first pair, or a new pair may be selected. At the next stage a new concept may be added to either of the existing groups, or the two groups may be joined together, or a third pair may be formed. At each stage the link added reflects the subject's judgement as to what represents the closest remaining relationship. The process is continued until all the concepts are linked to form a tree-like structure. Subjects may be asked to number the lines which constitute the graph consecutively as they are added (Shavelson 1974). In this case, a distance matrix is derived by adding the numbers associated with the line or lines joining each pair of concepts. A simpler procedure was used in the other studies reported above in which the separation between two concepts was taken as the number of links separating them in the tree. Preece (1976a) argued that this represents an over-estimate of inter-concept separations and, instead, proposed the square root of the number of separating links as a better measure.

All the five studies cited in the paragraph above, compared the structures derived by tree-graphing with those obtained using word association tests.[1] In a number of cases comparisons were also made with structures representing the subject matter involved (Shavelson and Stanton, 1975; Preece, 1976a; Fisher, 1979). Significant levels of convergence were reported in all cases, and Preece found that using his suggested square root separation measure improved such convergence.

---

1. The separation or distance matrices obtained by this method are often converted to proximity matrices. References to procedures for doing so are cited by Stewart (1979).

The three methods described above, all restrict the subject's response to the organisation of given concepts and enforce absolute decisions regarding their relative placings. This constraint has been regarded as a serious weakness by Preece (1976a) and by Fisher (1979). Tree-construction graphs have been fairly widely used in science education, and the reliability (Shavelson and Stanton, 1975) and the validity (Shavelson, 1974) of the method in the assessment of the cognitive structure of groups have been attested. However, Fisher (1979) states that the method lacks "interpretive power and validity at the individual level" and Preece (1976a) feels that the constraints imposed by the method probably make it "an unsuitable instrument for measuring inter-subject differences in concept interconnectedness".

5.2.5 **Word association tests.** Clinical interviews and digraphing methods attempt to interpret cognitive structure by means which require the subject, to a greater or lesser extent, to select and manipulate information within that structure. The constraints imposed on students' responses by the various concept-relating tasks described in the last section have already been referred to. By contrast, the word association test (WAT) is a method which interferes minimally with the structure it seeks to elucidate and which appears to put few constraints upon a subject's responses. Deese (1965) claims that the WAT is "perhaps as close to a context-free testing situation as can be devised". In these respects, and as regards the ease of collection of data, the WAT appears to offer clear advantages and has been the most widely used method for examining semantic relationships and structures in the context of science education (Preece 1976a).

In his pioneering work in seeking structural interpretations of WAT data, Deese (1962, 1965) was primarily concerned with the simplest form of WA testing. In this, each subject was presented with a series of stimulus words, one at a time, and was asked to respond to each with the first other word it brought to mind. From such tests he derived "associative meanings" for each stimulus word for a given group of subjects and calculated "intersection coefficients" reflecting the extent to which pairs of stimulus words shared responses in common. Further analysis of matrices of intersection coefficients led to the derivation of group "associative structures" for the selected set of stimulus words. Deese's approach, however, is based on the idea of associative structure as an attribute of a **population** and requires modification if it is to be applied to individual structural organisation.

Noble (1953) was amongst the earliest to make extensive use of **continued** WATs (in which students are required to make a number of different responses to each stimulus word, usually in a given time) as a tool to investigate the "meaningfulness" of particular words, whilst Deese (1965) pointed to the value of this technique in allowing "a possible extension of the method to the structure of an individal". A further distinction can be made between "free" and "controlled" WATs. In a controlled test, responses are restricted to a given category (for example, words used in science) and Preece (1976a) has reported that restricting responses in this way may sometimes be of value in mapping cognitive structure. The calculation of "relatedness coefficients" from the overlap between pairs of response lists (Garskoff and Houston, 1963) and the transformation of the resulting "relatedness matrices" into cognitive maps will be discussed in Sections 5.2.6 and 5.2.7.

The review of the use of WATs in research in science education which follows, will be of a summary nature and will be more concerned with the uses to which the technique has been put than with detailed findings, except as regards issues of immediate relevance to the present work.

A few studies have used WATs to explore the meanings which students attach to particular science concepts. Of most interest is the work of Schaefer on the concepts of "growth" (Schaefer, 1979a) and "ecosystem" (Schaefer, 1979b). He described a structural model in which a "logic core", investigated by asking students for a free definition of the concept, is embedded in an associative framework tapped by free continued WA testing. In a cross-sectional study of the development of the concept "ecosystem" he found that older students and teachers tended to have narrower and more abstract concepts. He also reported a correlation "in excess of 0.7" (N = 421) between the distribution of free associations and that of free definition elements (Schaefer, 1979b). Schaefer's work has been carried out in West Germany, Israel and the Philippines, and other cross-cultural studies of the associative structures of students' science concepts have been reported by Chao (1974), Isa and Maskill (1982) and Ross and Sutton (1982).

A number of studies have related data from WATs to students' performance in science. In work with high school physics students, using an inter-related group of physics (mechanics) concepts, Johnson (1965) found a significant correlation between the number of a student's responses which were also words in the stimulus list, and his performance on a related problem-solving test. In a subsequent study he also found high achievement to be associated with the total number of associations generated by a student and with the number of stimulus-list responses in the top halves of his response lists (Johnson 1967). Shavelson (1973) reports a similar finding in the course of an experimental study in which the number of such responses in the upper half of a student's list proved to be a good predictor of post-test achievement. More able students have also been reported to use more "scientific" and more abstract associations (Maruping 1980). Recent reports by Kempa (1982) and by Kempa and Nicholls (1983) relate to comparisons between the associative maps of groups of O-level chemistry students identified as good and poor problem-solvers. This work was contemporary with that of the present study; it was carried out in cooperation with the author and will be referred to again later. Good problem-solvers were found to have more complex associative maps with higher levels of relatedness between key concepts. The maps of poor problem-solvers were deficient with respect to links involving more abstract concepts. Similar findings have also been reported by Moynihan (1982).

A number of cross-sectional studies have used WATs to trace the differences of cognitive structure associated with different levels of education. The work of Schaefer (1979b) on the structure surrounding a single concept has already been referred to. Johnson (1964) used 18 inter-related mechanics concepts in single response WATs with high school students who (i) were not planning to study physics, (ii) were planning to study physics, (iii) were studying a physics course, and (iv) had recently studied physics. He reported that certain associative measures, and in particular the number of stimulus-list responses, were related to a student's degree of involvement in physics. In a similar study, Preece (1976b) used correlations of the rank orders of relatedness coefficients to demonstrate that agreement with three alternative subject matter structures increased with levels of physics education. A growth of associative meaning tending from the concrete to the abstract, has also been demonstrated for concepts in chemistry between Form I and Form VI (Wijenayake 1978).

Changes in cognitive structure following instruction have been investigated using word association methods, though not all have been truly experimental studies using control groups. Thus Johnson (1967) reported a significant correlation between the WA responses of high achievers and the co-occurrence of words in a physics text which they had studied, and Rothkopf and Thurner (1970) found that students' word association patterns before and after studying a physics text showed changes in the direction of word patterns in the text. In an important controlled, experimental study Shavelson (1972, 1973) investigated the correspondence between students' cognitive structures (as assessed using a WAT) and the organisation of the same concepts in a physics text (as assessed by digraph analysis) before and after they had studied it. He found that, following instruction, the experimental group produced more responses than the control group and tended to use more words that defined and controlled the stimulus concepts. They also showed increased inter-relatedness coefficients and significant convergence with the subject matter structure in the text, while the control group did not. It may, however, be noted that a study by Dangerfield (1977) using digraph methods only, suggested that students did **not** tend to acquire cognitive structures reflecting the text which they had studied when the latter was organised in an unusual way.

A number of WAT studies relate to science learning experiences other than the study of texts. In an experimental study with seventh-grade general science students, Morris (1970) found changes in WAT response lists in the expected direction and a rise in mean relatedness coefficients following instruction. Similar findings have been reported by Cachapuz (1979), Cochaud (1979), and by Moreira and Santos (1981) who also refer to qualitative changes in cognitive maps in the direction of more logical structures. An experimental study by Thro (1978) found that students subjected to physics instruction showed changes in associative structures towards convergence with that of the instructor. Finally, an interesting study linked with the work of Ausubel has been reported by Gunstone (1981). Two instructional programmes, with differing emphases on linking to existing knowledge and experience, were used to teach elementary dynamics to physics students. Outcomes were evaluated using a modified WA technique to assess linking in cognitive structure between dynamics and existing knowledge. It was found that differences in performance between the two groups on far-transfer problems were mediated by the extent to which the new learning was linked to existing cognitive structure.

The focus of another group of studies using WAT, has been to evaluate the technique itself as an approach to the mapping of semantic structures in science. All those reviewed have reported satisfactory levels of convergence between structures based on WAT data and those derived either from alternative proximity measures or from an analysis of the subject matter or both. Ten such studies have been listed in Table 5.1. Johnson, Cox and Curran (1970, 1971) concluded that similarity ratings and WATs access the same underlying structure and Shavelson and Stanton (1975) present convergent data which support the interpretation that WATs measure "a significant part of cognitive structure". Preece (1976a) concluded that a WAT is a valid means of mapping cognitive structure. He also found that the results obtained from a controlled WAT conformed more closely to a model of subject matter structure than did those from a free WAT.

> ".... the word association test was a particularly valuable device for mapping cognitive structure if associations were constrained to the particular semantic domain involved."
>
> (Preece 1976a, p 7)

Table 5.1 Studies evaluating word association tests as a means of elucidating semantic structure

Citation	Subject studied	No of concepts	Comparison made
Johnson, Cox and Curran, 1970	Mechanics	6	WAT and similarity ratings.
Johnson, Cox and Curran, 1971	Mechanics	9	Free and controlled WAT and similarity ratings.
Shavelson, 1972	Mechanics	14	WAT and subject matter structure.
Shavelson, 1974	Operational systems	12	WAT and tree-graph construction.
Shavelson and Stanton, 1975	Operational systems	12	WAT, tree-graph, card sorting and subject structure.
Preece, 1976a	Mechanics	15	Free and controlled WAT and tree-graph construction.
Preece, 1976c	Electro-magnetism	15	WAT and subject matter structure.
Johnson, 1976	Reaction kinetics	12	WAT and digraph of students' essays.
Cachapuz, 1979	Reaction kinetics	14	WAT and similarity rating.
Fisher, 1979	Relative humidity	16	WAT, tree-graph, a laboratory "stations" method and subject matter structure.

Finally, some interesting non-structural interpretations of data from WATs used in a cross-sectional study have been reported by Preece (1977). The tests, consisting of 15 mechanics concepts, were administered to groups of physics students in their first, fourth and seventh years of secondary school and to a similar group of science graduates. Two non-physics groups (seventh year and university) were also tested. Four indices were derived and the development trends in each were examined. An index M represented a student's Mean number of associations per concept, and increased with age irrespective of whether or not the individual continued to study physics. D represented a student's mean number of Directly related responses, defined as words which occurred with the stimulus word in one of the fundamental equations of mechanics. For physics students D rose steadily up to the seventh year of secondary education after which it remained fairly constant. For non-physics students it declined with the time elapsed since the student last studied the subject. U represented a student's mean number of Unrelated stimulus-list responses; it tended to increase quite rapidly up to the fourth year but fell away after that. Finally R represented a student's mean Relatedness coefficient. This showed a similar trend to U although falling away more slowly. These results are consistent with the hypothesis that, as learning starts, the total number of associations for a new concept increases rapidly. Later, discrimination occurs between directly related associations and the rest. Some confirmatory evidence for this hypothesis has since been reported (Preece 1978a).

5.2.6 **The calculation of relatedness matrices from WAT data.** The derivation of a map representing cognitive structure from WAT data involves two stages, namely (i) the generation of a half-matrix of "relatedness coefficients" which reflects the associative relationships between all pairs of stimulus words, and (ii) the interpretation of this half-matrix so as to identify the underlying organisation of concepts in the learner's mind. The first stage of this process is considered below.

Deese (1962) pointed out that the "associative meanings" of two words are similar to the extent that they stimulate responses in common and goes on to describe a method in which such common responses are used to generate "overlap coefficients". Deese's analysis applies strictly only to single-response WAT data. However, it has already been argued that continued association is the appropriate tool when individual differences are of interest. Cofer (1958) was able to show that ".... discrete single word association and continued association produce roughly comparable results". Likewise, Laffal and Feldman (1962) demonstrated that matrices based on single and continued associations showed similar structures. Garskoff and Houston (1963) have developed a method of analysing continued associations which is consistent with the concept of cognitive structure which the present study shares with their work, and Preece (1976a) has described this as the most widely used technique in the field. The method is briefly described below.

A completed continuous WAT for an individual student consists of a set of stimulus words, to each of which is appended a list of response words. For a pair of stimulus words, a relatedness coefficient (RC) is calculated using the formula:

$$RC = \frac{\overline{A} \cdot \overline{B}}{(A \cdot B) - [n^P - (n - 1)^P]^2}$$

where:

- A refers to a list of m words (a_1, a_2 .... a_j .... a_m);

- B refers to a list of n words (b_1, b_2 .... b_i .... b_n), and $n > m$;

- $\bar{A} \cdot \bar{B}$ is a (weighted) measure of the overlap between the two lists calculated by (i) assigning "scores" n^p, $(n - 1)^p$ .... 1^p to successive words in lists A and B (terminating in the case of the shorter list with the m^{th} term), then (ii) for each pair of words a_i, b_j common to the two lists, multiplying the assigned scores, and finally (iii) summing all such products;

- p is a (weighting) integer whose value is discussed below;

- $$(A \cdot B) = (n^p, (n-1)^p, \ldots\ 1^p) \begin{pmatrix} n^p \\ (n-1)^p \\ \vdots \\ 1^p \end{pmatrix} = [n^p]^2 + [(n-1)^p]^2 \ldots + 1.$$

The derivation of this formula is discussed in detail by Garskoff and Houston (1963) but a few points will be reviewed here.

In Garskoff and Houston's formula, a_1 and b_1 refer to a pair of stimulus words, the remainder of the lists consisting of an individual's ordered responses. Thus, in calculating the measure of overlap represented by the numerator $\bar{A} \cdot \bar{B}$, each stimulus word is regarded as its own first response. This so called "representational response" is an essential assumption not only of this, but also of Deese's (1962) original method. It is intuitively reasonable to assume that a person's first response to a word as stimulus is to identify or recognise the word and locate it in LTM prior to seeking associates, and indeed subjects are generally instructed to say the word to themselves before responding. The representational response is also felt to be a necessary assumption to deal with the not uncommon case of a pair of words which, whilst having no overt responses in common, frequently elicit one another as responses. Such pairs are obviously associatively related and the representational response provides a device allowing this to appear in the computed coefficient. For a more detailed discussion, see Deese (1965, pp 46-47).

The numerator in the above equation represents a measure of the overlap occurring between two different stimulus words, whilst the denominator represents the maximum possible overlap between them. (The term A . B represents the overlap between identical lists, while the "correction term", subtracted from this, makes allowance for the fact that, for **different** stimulus words, the maximum overlap occurs where each word is the first response to the other, so that the first two words in the list are reciprocally related rather than identical). The coefficient RC can thus take a value between zero (no associative overlap) and unity (a "perfect" overlap, with each stimulus word the first response to the other, the remainder of the lists being identical).

In considering the implications of the construct "cognitive structure" for its measurement using WA testing, Shavelson and Stanton (1975) emphasised that:

> "The order of concepts retrieved from LTM takes on a particular importance in making inferences about the organisation of concepts in LTM."
>
> (Shavelson and Stanton 1975, p 73)

The integer p allows different weightings to be attached to this "response primacy". The value of p has usually been taken as one, resulting in a linear weighting, the first response being evaluated as n times more important than the n^{th}. However, other values have also been employed. Preece (1976a) compared the effect of using p = 0 (no weighting for primacy) and p = 1, as an element in a comparative study of methods for mapping cognitive structure and concluded that:

> ".... giving less weight to later, presumably more idiosyncratic, responses resulted in a better mapping of the common cultural aspects of the cognitive structure as represented by a theoretical structure."
>
> (Preece 1976a, p 7)

The authors of the method, Garskoff and Houston (1963), suggest that stimulus words typically having a few dominant associates may be thought of as having steeply sloped response hierarchies and thus as demanding a relatively high value of the exponent p as compared to words with flatter response hierarchies. They also propose that personality variables, such as creativity, might be expected to correlate with a tendency to produce steep or flat response hierarchies, thus demanding different weighting exponents. They go on to describe an empirical study suggesting "that two is the appropriate value for p for low creatives while one is more suitable for highly creative subjects". Thus, while a value of one for p has been an almost universal choice in studies of cognitive structure in the field of science education, consideration might also be given to assigning p a value of two in view of (i) the steep hierarchies which would seem likely for words representing science concepts, (ii) the tendency for less creative individuals to be directed towards science studies and (iii) earlier findings suggesting that the upper parts of response hierarchies correlate most highly with PS achievement (Shavelson, 1972).

Finally, it has been normal practice to calculate mean relatedness matrices for groups of subjects, in which every cell entry is calculated as the mean of the RCs in the corresponding cells in the individual matrices. The concept of mean relatedness coefficients is not objectionable as the measure has more than merely ordinal significance (a difficulty with certain psychological "distance" measures). Thus, given the assumption implied by a particular weighting coefficient p, it is correct to regard a coefficient of, say, 0.6 as representing twice the associative overlap as that represented by a coefficient of 0.3. Thus, for example, if p = 0 the first pair of words shows twice as many responses in common as the second.

5.2.7 The transformation of a relatness matrix into a cognitive map. A number of methods have been described for transforming matrices of relatedness coefficients, which seek to organise and interpret the data in ways which can be regarded as mapping corresponding cognitive (semantic) structures. Preece (1978b), in a recent review of such methods, refers to the use of (i) spatial maps resulting from a technique known as "multi-dimensional scaling" (Shepard, 1962; Kruskal, 1964), (ii) nested hierarchical maps which are produced by a number of cluster analysis methods, and (iii) graphical maps which relate to a network method developed by Waern (1972). In order to accommodate factor analysis and a non-hierarchical clustering method (Jardine and Sibson, 1968), these headings have been slightly modified in the comparative account which follows. Attention will be focused mainly on the nature of the solutions offered by each method, rather than on the findings of particular studies.

(i) **Dimensional methods.** Dimensional methods include factor analysis and multi-dimensional scaling, both of which employ mathematical models to reduce an original n-dimensional[1] matrix to some optimally meaningful lesser number of dimensions.

While dimensional methods have been quite widely used, the most successful cases appear to have been those in which a particular dimensional interpretation might have been anticipated in advance. For

1. For a matrix derived from WAT data n represents the number of stimulus words used in the test.

example, a series of studies with verbal auxiliaries, personal pronouns and conjunctions (discussed in Deese, 1965) reproduced recognisable grammatical classes when the relatedness matrices were factor analysed, and Shavelson (1974) has reported a study by Fillenbaum and Rapoport in which multi-dimensional scaling of data based on WA responses to colour words reproduced the familiar colour circle in two dimensions. In this respect it is relevant that Michon (1972), in a paper comparing dimensional with hierarchical clustering methods, selected examples of the former in which the data clearly lent themselves to dimensional interpretation (for example a city mapping exercise and a task involving the memorisation of the relative position of pairs of elements in a square matrix). Where no a priori expectation of this nature exists the results of such studies often seem less impressive. After reviewing a number of dimensional studies, Waern (1972) concluded that a simple clustering approach "may provide a valuable alternative to dimensional methods especially in cases where the cognitive structure is a priori quite unknown". She suggested that there may be cases where "cognitive structure may best be described by local associations and not by global dimensions". Indeed Dietze (1963), in a comparison of analytical methods, has questioned the applicability of factor methods to data matrices composed of indices other than correlation coefficients. After an empirical investigation he concluded that a simple clustering method offered many advantages and that the "types" yielded were comparable with, and possibly superior to, those yielded by factor methods. Finally, Michon (1972) has noted that not only are dimensional and clustering methods compelementary, but that under certain mathematical conditions they can be shown to give identical results. It thus appears that clustering techniques need not obscure a possible dimensional interpretation, while at the same time offering greater flexibility.

The use of multi-dimensional scaling for mapping cognitive structures in science has been well reviewed by Preece (1978b). A number of the studies reported yielded acceptable two dimensional solutions, and it seems possible that hierarchical or graphical solutions may produce equally valid, and more readily interpretable, results. In this connection it is also of interest that, although no explanation is given, Shavelson, in his series of studies involving cognitive mapping, shifted from the use of multi-dimensional scaling (Shavelson, 1972) to hierarchical cluster analysis (Shavelson, 1974).

(ii) **Hierarchical clustering methods.** A large number of clustering methods have been discussed comparatively by Everitt (1974) of which the simplest ("nearest neighbour single linkage") seems to have been the most widely used. This was the method selected for example by Shavelson (1974) and Shavelson and Stanton (1975) and attributed by these authors to S C Johnson (1967) although the latter author admits that the method had been previously described on more than one occasion. The method is most readily described in terms of a simple algorithm thus:

1. the matrix is searched for the pair(s) of most similar entities (ie for the largest coefficients);

2. the two rows and columns so identified are collapsed by taking the <u>highest</u> values of corresponding pairs of cells;

3. the process is repeated until the matrix is reduced completely.

At each stage the entries identified are added as new terminal nodes on a horizontal axis, their relationship to other entities being indicated by constructing an hierarchical tree-graph in which the vertical axis represents, usually to scale, the values of the coefficients corresponding to each non-terminal node (ie to each conjunction of branches in the tree).

S C Johnson (1967) showed that this hierarchical analysis assumes that the matrix of coefficients satisfies the ultrametric inequality,[1] although Michon (1972) found that the method yielded "more interpretable" results than dimensional methods even when this condition was violated. Jardine and Sibson (1968), dealing with the mathematical implications of hierarchical clustering techniques, noted that the ultrametric constraint can rarely be met and proposed criteria which any attempt to impose an hierarchical solution should meet in such circumstances. Of methods described, only the nearest neighbour approach meets all these criteria, however it has the disadvantage of tending to produce chains, rather than clusters of entities. As Jardine and Sibson (1968) have pointed out, this is a feature of the analysis rather than a fault, nevertheless the chains may obscure information that is of interest. An "average linkage" method which attempts to overcome this problem will therefore be considered next.

In the nearest neighbour method, entities are assigned upon the basis of their closest relationship only and this may give rise to distortions. To take a hypothetical case, ELECTRON might link to a cluster of concepts relating to, say, electrolysis because its highest relatedness coefficient happens to be with NEGATIVE, which in turn is linked to CATHODE. However, examination of the relationship of ELECTRON with other concepts might indicate a whole set of only marginally lower coefficients suggesting it should cluster with several concepts relating to atomic structure. This information, which might be of great interest, would be lost in the nearest neighbour analysis. Many similar

1. $d(x, y) \leqslant \max\{d(x, z), d(y, z)\}$ where $d(x, y)$ corresponds to a measure of the distance separating x and y etc; the proximity measures, p, in a relatedness matrix can be simply transformed to a corresponding distance measure using a formula such as $d(x, y) = [1 - p(x, y)]$

attempts to overcome this sort of problem have been reported, of which one described by McQuitty (1966) is representative. In this, after initial fusions of pairs of concepts, columns and rows are collapsed on the basis of **average** rather than highest relationships. This method fails to meet only one of Jardine and Sibson's (1968) criteria, namely that of "continuity" (so that progressive small changes in the data may not necessarily produce corresponding small changes in the resulting hierarchical dendograms) and may well produce a more satisfying interpretation of WAT data than the nearest neighbour method.

Hierarchical methods have not been very widely used in mapping cognitive structures in science. Shavelson (1974) and Shavelson and Stanton (1975) report work in a highly structured topic area (operational systems) in which the nearest neighbour method gave readily interpretable results. Moreira and Santos (1981) also felt able to offer interpretations of maps obtained by the same method in the field of thermodynamics. However, the present author feels that the dominance of chaining which is shown in the published diagrams may render such interpretations of limited value.

(iii) **Non-hierarchical solutions.** A non-hierarchical clustering method, devised by Jardine and Sibson (1968), overcomes the technical problem of the ultrametric by finding **overlapping** clusters. The method has been employed in linguistic analysis and the overlapping feature is consistent with the model of semantic structure underlying the use of WATs. The fact that a concept may appear in two (or more) clusters solves the sort of difficulty referred to in the hypothetical example quoted earlier involving the concept ELECTRON. Using Jardine and Sibson's method, ELECTRON could belong to both the "electrolysis" and the "atomic structure" clusters.

The method seeks groups of entities (called maximal complete sub-groups) which are fully interconnected above a given threshold coefficient, that is to say, sub-sets of concepts for which all pairs are connected above the threshold level. Next, all such sub-groups intersecting in at least K (a positive integer greater than 1) points are merged, resulting in a set of clusters which may intersect in up to (K - 1) points. Unlike the nearest neighbour method, which tends to yield chains, and the McQuitty (1966) method which tends to yield spherical clusters, this approach imposes no constraints on the shapes of clusters. However, while the two former methods automatically lead to dendograms (hierarchical tree-graphs), Jardine and Sibson's method requires decisions regarding the threshold level selected and the value of K to be employed. The results are probably best presented in the form of Venn diagrams. The method is considerably more demanding on computer time than the other approaches referred to and does not appear to have been applied to the mapping of cognitive structure in science.

A very simple graphical method of processing a matrix of relatedness coefficients has been described by Waern (1972). The method makes few assumptions, allows both clusters and chains to emerge, and permits a single item to be linked to more than one group. On the basis of empirical evidence, which she reports for quite a wide range of examples, she claims that her approach yields results which have similar, and in some cases superior, interpretive power to those obtained by much more elaborate dimensional methods. She also regards the method as superior to most clustering techniques in that both clustering and chaining are allowed, and there is no insistence that each entity must belong exclusively to a single cluster.

Waern suggests two variations of her method. In the single-step method, a threshold relatness coefficient is selected and a simple graph is constructed in which all pairs of concepts (each concept is represented by a point) which are related at or above the threshold are joined by lines. The cut-off point is chosen so as to (i) minimise the number of isolates (unconnected points on the graph) whilst (ii) maximising the distance between the two most distant points on the graph. These two criteria act in opposite directions and the balance between the two is a matter of judgement. The preferred multi-step method is a development of this procedure which is simply repeated using a series of decreasing thresholds until no further useful information is obtained. At each step new connections already accounted for by paths previously obtained, are not included in the graph. This prevents the collapse of chains into clusters. Waern's method has been used in a number of studies for mapping cognitive structures in science and Preece (1976a) has claimed that a great advantage of the multi-step method is the extra clarity afforded by the initially very simple structures out of which the final structure grows. He obtained maps which showed, "perhaps even more clearly" than those derived in other ways, interpretable structures for both mechanics and electrostatics concepts (Preece 1976a, b, c). The method has also been used successfully by Cachapuz (1979) in the field of reaction kinetics and by Fisher (1979) with concepts relating to relative humidity. Since this report was drafted (prior to carrying out the second phase of the study early in 1980), further uses of Waern's technique in the field of elementary chemistry have been reported by Kempa (1982) and by Moynihan (1982).

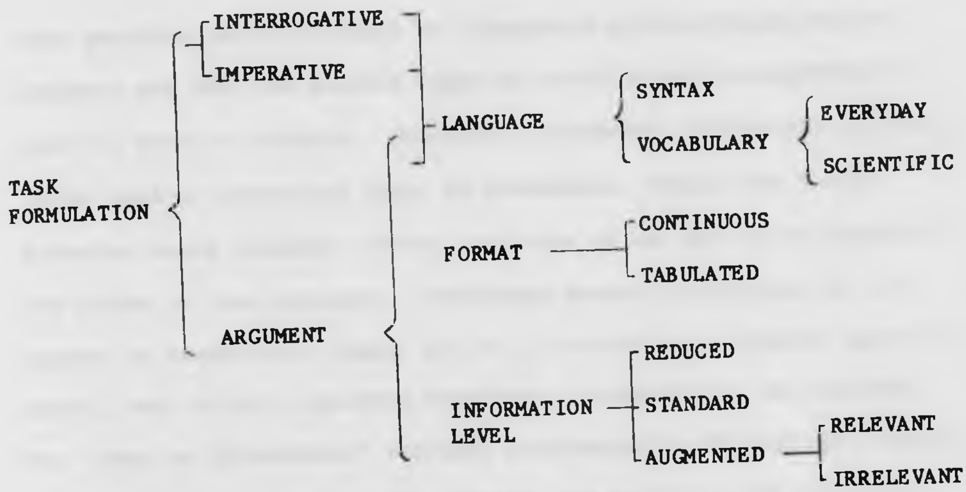
5.3 The formulation of examination tasks

A student's task in solving an examination item may be broadly defined by the answer demanded of him. Thus the nature of the goal was used in classifying tasks during the construction of the test employed in the first phase of this study. However, it seems likely that the way in which a task is presented may influence the way in which a student tackles it. The choice of a particular word or format may, for a given student, provide the essential cue to the required knowledge structure in LTM or, indeed, may throw him off on the wrong track. For example, Worsley (1979), referring to work carried out by the Assessment of Performance Unit, reported that the mere addition of a few numbers to one of their science items, without altering the task to be performed in any way, was sufficient to depress the facility of the item quite significantly, and also that changes in the context in which the task was presented appeared to influence performance. A systematic study of the way in which selected task formulation variables influence performance might be expected to shed further light on students' PS behaviours and to generate information of interest regarding the design of examination items.

An analysis of task formulation in the context of short-answer, non-mathematical items, is presented below and the selection of a particular task formulation variable for further study is discussed.

5.3.1 **An analysis of task formulation.** Passing reference has been made by Ogborn (1979) to attempts to develop networks for describing examination items in terms of their content and the skills needed to solve them. The conventions used in constructing such networks were described in Section 4.2.5 and have been applied to the question of task formulation in Figure 5.1.

The network can be quickly described. The task represented by an examination item consists of either an **interrogative** ("What is the



Conventions employed: { represents a logical AND;
[and] represent a logical OR

Figure 5.1 A Network Analysis of Task Formulation[1]

name of ....?") or an **imperative** ("Explain what is meant by .....") together with an **argument** which refers to a particular chemical situation. The argument may vary in length from a word or two to several sentences. Three variables are used to describe the argument: language, format and information level. **Language**, which also describes the interrogative or imperative part of the item, is seen in terms of syntax and vocabulary, the latter consisting of everyday words and chemical (and other science) terms. The **format** of an item refers explicitly to whether or not tabulated information is presented. Finally, the idea of **information level** is introduced to denote the extent to which the amount of information provided in the argument, matches the demands of the task. In a "standard" examination item the match is a good one; the amount of information provided is sufficient, and only sufficient, to enable the problem to be solved. However, it is

1. The network was developed to describe short-answer, non-mathematical chemistry items; however, it could readily be adapted to accommodate a much wider range of tasks.

also possible that the amount of information provided might not be adequate and that the student might be forced to make assumptions in order to reach a solution. Conversely, redundant information (relevant to the task or otherwise) might be introduced. Whilst the former situation would probably reflect a mistake on the part of an examiner, the latter is less uncommon. Irrelevant textual information is not unusual in examination items, nor is it unusual for a limited amount of useful, but strictly speaking redundant, information to be included. The "level of information" variable is principally of interest, however, in considering the relationship between the artificial PS situation encountered in examinations, and PS in a wider context. Outside the examination hall the information available is not always perfectly matched to the task in hand, and the problem-solver may have to make decisions regarding both its adequacy and its relevance. Task formulation variables are further discussed in the next section.

5.3.2 **A discussion of task formulation variables.** The influence of the language used in constructing multiple-choice items, on candidates' performance in chemistry at O-level, has been studied by Cassels and Johnstone (1977). They considered both syntax and vocabulary and found, particularly for students of low verbal ability but to some extent for all candidates, that item facilities were significantly influenced by the use of (i) scientific terminology in key positions, (ii) unnecessary verbosity, (iii) negatives, (iv) ambiguity, and (v) the way in which clauses or sentences were put together. They also confirmed Gardner's (1972) findings that many of the non-technical words habitually used in science teaching and examining are not well understood by students. Further work has continued to document students' difficulties with a large number of non-technical words and has also shown that some words, which are well understood in everyday situations, may cause problems when used in a scientific context or with a scientific connotation (Cassels and Johnstone, 1980, 1983).

The term "format" has been used in the network presented in Figure 5.1, to distinguish between items in which all the information is presented more or less continuously in normal prose, and those in which key information is tabulated. No published work bearing directly on the influence which item format (in this sense) may have on a student's PS behaviour could be identified during the literature search undertaken for the present study, although two of the processing errors reported in the first phase of the study involved mistakes in handling tabulated data (see accounts relating to errors E2f and E2d in Section 4.5.3). The incidence of such errors was, however, very low and it was felt that no priority could be given to further consideration of this variable.

"Information level" was considered to be of potential interest as a task variable from two points of view. Firstly, it was felt that it might be useful as a tool for the further exploration of students' PS behaviour, and secondly that it might provide a means of bridging the gap between idealised examination tasks and PS in a wider context. As regards the former, it is relevant to consider the work of Krutetskii (1976) who designed sets of mathematics items with incomplete and with redundant information, to study students' "perceptions of relations and concrete facts in a problem". In a series of one-to-one interviews in which students tackled problems aloud and were questioned about their reasoning, he found that the items discriminated very clearly between mathematically gifted and average children. While the former quickly perceived the structure of the problem and were thus able to identify deficiencies or redundancies in the information provided, average and below average students tended to process items of information one by one in isolation. They were unable to recognise that items with missing information were insoluble, or to specify what was missing when this was pointed out to them, and they were confused by redundant information which interfered with their solution of the problems. This work obviously brings to mind Gestaltist views regarding the importance of structure in PS. In particular, Wertheimer (1959) has reported that students taught by a method emphasising the structural properties of geometrical figures were able to identify certain classes of insoluble problems in which insufficient information was provided, while those taught by traditional methods were not.

As regards the relationship between PS in examinations and elsewhere, a recent paper by Ward (1982) relates to chemistry students' problems with the interpretation of observational data in the context of A-level inorganic analysis. He considered the uncertain quality of the observations likely to be made by students and identified types of "observational data sets" which correspond closely to the information levels described in Figure 5.1. He found that, while some students recognised the limited nature of the conclusions which could be drawn from incomplete information, others tended to make unjustifiable assumptions. And while irrelevant data were largely ignored, attempts were usually made to process redundant information of a relevant kind although this added nothing to the solution. Like Krutetskii, he found that most students tended to treat items of information in isolation from one another, rather than integrating them into a coherent analytical scheme.

5.4 <u>The aims and methods of the second phase of the study</u>

The broad aims of the second phase of the study have already been referrred to. They were:

(i) to explore differences in relevant aspects of the cognitive structures of more and less successful examination candidates, relating these to PS behaviour, and

(ii) to study the influence on candidates' PS performance of the level of information provided in an item.

These aims are discussed below and a brief account is given of the methods adopted in the study.[1]

5.4.1 **The aims of the study.** On the basis of the findings of the first phase of the study, as reported in Chapter 4, it was hypothesised that successful candidates differed from less successful principally in terms of the organisation of chemical knowledge in LTM. It has also been suggested, in the present Chapter, that examination performance tends to be related to subject matter content. It was therefore predicted that students who performed well on items relating to a particular topic area could be shown to possess more systematically organised semantic structures, in respect of concepts in that topic area, than those who performed badly on the same items. More explicitly, that their structures would show higher levels of linkage overall and superior discrimination between more and less closely related concepts. In addition to testing this rather broad hypothesis, it was hoped that it would be possible (a) to improve general interpretations of students' strategies, and in particular errors, in terms of their cognitive maps, and (b) to interpret particular PS episodes of successful and unsuccessful students in terms of individual cognitive maps.

Three major task formulation variables have been identified as language,

1. It may be relevant to record that this work was planned in the latter part of 1979 and carried out early in 1980. Although in the discussion in the early part of this chapter, relevant citations have been brought up to date (to mid-1983), the planning of the work reported here was obviously not influenced by any materials appearing after the early part of 1979.

format and information level. Of these, the role of language in examination items in chemistry is already under investigation (Cassels and Johnstone, 1983) and item format was considered to be of relatively limited interest. However, broad indications have been reported that many students may have difficulties coping with uncertainties regarding the match between levels of available information and the demands of the task in hand (Krutetskii, 1976; Ward, 1982). A study of the influence of levels of information on PS behaviour was therefore considered to be of interest in attempting to bridge the gap between PS in examinations and PS in more realistic chemical situations where levels of information are less well controlled. Given a particular internalised map of chemical knowledge, it might be anticipated that the likelihood of success at any particular task would increase with the level of information supplied. Thus, reducing the level of information in an item should reduce the number of potential cues available for activating a search in LTM; similarly, adding relevant information should increase the number of potential cues. A well integrated knowledge structure might be relatively insensitive to such cuing levels insofar as the whole structure would become available on cuing any part. However, the performance of students with more fragmented knowledge-bases would be expected to improve as the number of relevant cues was increased. It was therefore hypothesised that the facility of an item would tend to reflect changing levels of relevant information, but that such effects would be less marked for high achievers. In terms of the analysis of task formulation presented in Figure 5.1, it was predicted that, particularly for average and below average students, facilities would tend to increase along a series of formulations of the same task using reduced, standard and augmented (relevant) levels of information.

However, a second factor also had to be considered in respect of items with reduced levels of information. When relevant information is removed from a normal examination item a degree of ambiguity generally results (see Section 5.5.1, part (ii), below). This was not expected to influence the trend already predicted insofar as most students were not expected to recognise such ambiguities. However, the effect on high achieving students was harder to predict. It was anticipated that they would be more likely to notice any ambiguity, but equally that they would be better able to make any necessary assumptions. Their performance might thus depend upon their attitude to encountering ambiguity in an examination context as well as upon purely cognitive factors. It was hoped that, providing reduced versions of items could be written in such a way as to demand only fairly obvious and neutral assumptions, the influence of such personality variables could be kept to a minimum so that the predictions already made would continue to apply.

As regards the influence of augmenting an item with irrelevant information, it was predicted that this would tend to depress item facility. In the studies already referred to, both Krutetskii and Ward reported that their subjects tended to process given information piece-meal. Since Krutetskii suggests that only the gifted student readily discriminates essential from inessential information, this is likely to lead to the tapping of the LTM knowledge-base at several points. In cases where the semantic relationships in that knowledge-base are not well discriminated, the result is likely to be confusion over selecting the most appropriate response and a reduced likelihood of success. Higher achievers would be less influenced by such effects insofar as (a) they may be better able to discriminate the demands of the task and thus to ignore irrelevant or misleading cues, and (b) they may possess better discriminated cognitive structures and thus be better able to select the most appropriate response in situations where they may not have succeeded in eliminating the effects of excessive cuing.

The aim of this part of the study was thus to examine the influence on performance of systematic changes in the level of information provided in an item, to interpret any systematic effects in terms of an information processing model of PS behaviour (and of relevant aspects of cognitive maps if practicable) and to relate any findings to realistic PS situations in which levels of information may not necessarily be well matched to the task in hand.

5.4.2 **The general methods adopted in the study.** It will be convenient to record first the straightforward methodology employed in the task-formulation aspect of the study. An analysis-of-variance design was used to compare the performance of groups of students assigned at random to different versions of the same task. The writing of the different item versions, their compilation into four tests and the administration of these to O-level candidates is described in Section 5.5 below. The analyses of variance and their results are detailed in Chapter 6. It had been hoped that supplementary information concerning students' PS strategies might be obtained using a "self-reporting" form alongside the tests. However, attempts to develop such a form, which are described in the next section, met with very limited success.

The selection of continued word association testing as the method employed in mapping cognitive structure presented little difficulty. It has been the method of choice in many studies in science eduation; it has, in particular, the advantage of being relatively context-free as compared to alternative methods and is simple and rapid to administer to large groups of subjects. Before proceeding to a description of the compilation and administration of matched PS and WA tests, however, it is necessary to discuss the reliability and validity of WATs and to consider recent published criticism of the method.

The reliability of continued WA testing was firmly established by Noble (1952) in the context of his study of the relative "meaningfulness" of a wide range of words. In the field of science education, similar findings in terms of the stability of the derived associative structures (in the absence of further relevant experience/learning) have been reported by Johnson (1967), Shavelson (1972, control group data) and Preece (1976a, b). The latter author used a test/re-test model and correlated the rank orders of the resulting overlap coefficients yield-reliabilities for various groups of students in the range 0.81 to 0.92. Later work by the same author (Preece, 1977) showed similarly good reliabilities for indices derived from controlled WATs (0.70 to 0.93).

The question of the validity of employing relatedness coefficients derived from continued WATs as a basis for mapping cognitive structure has been reviewed by Shavelson (1974), Shavelson and Stanton (1975) and Preece (1976a). This validity rests upon (i) the convergence of a number of different methods of establishing cognitive maps (concurrent validity) and (ii) the fact that agreement between cognitive maps established from WAT data and the structure of the underlying subject matter have been consistently found to improve with increasing levels of relevant instruction; this may be construed as construct validity (as defined, for example, by Lewis (1973)) insofar as WAT-derived maps successfully distinguish those who have been more exposed from those who have been less exposed, to a given subject matter structure. Eight examples of the convergence of different methods in the science field were presented in Table 5.1. Shavelson and Stanton's (1975) study using simple card-sorting, tree-graph construction and a WAT with a topic from mathematics was explicitly concerned with establishing concurrent validity and their findings are typical. On the basis of generating and comparing hierarchical maps they concluded that the three methods "measured principally the same construct". They also compared the associative maps of experts with a representation of the subject matter structure and concluded that this method assessed "a significant part of cognitive structure". Other findings which provide evidence of construct validity include those reported by Shavelson (1972, 1974) and by Preece (1976c). Referring to experimental work in which students were given WATs following instruction every day throughout a six day physics course, Shavelson (1974) writes:

> "If these data [ie data derived from the WATs] measure some important aspect of cognitive structure, then the cognitive structure of students receiving instruction should resemble the subject-matter more closely with each day of instruction, while the cognitive structure of students not receiving instruction should be no more similar to subject-matter structure at the end of testing than at the beginning."
>
> (Shavelson 1974, p 241)

The expected trend was clearly demonstrated. Indeed, insofar as the studies reviewed above have found positive relationships between data obtained from WATs and (i) relevant cognitive performances, (ii) levels of relevant education, and (iii) explicit relevant learning experiences, such studies tend to contribute to the argument for the construct validity of WAT-based measures of cognitive structure.

Evidence has been presented above of the reliability and validity of WATs as a basis for mapping cognitive structures in science. Attention has also been drawn, in the literature review in Section 5.2.5, to many cases of their successful application for this purpose. However, criticisms of the methods of "associative mappers" have been advanced by Stewart (1979) on the grounds that the inference of structure from associative data is unsound. He also complains that maps based on semantic proximity data fail to represent the propositional nature of information in cognitive structure; that they ignore the **nature** of the associations underlying such data and fail to distinguish those reflecting correct, from those reflecting incorrect, propositions. Sutton (1980) and others concerned to elucidate children's "alternative frameworks", have echoed the latter criticism and Sutton also refers to an element of randomness in free associations.

Stewart's paper has recently been refuted by Nagy (1983), who considers the inference of structure from WAT data no less sound than that implied in the use of classroom tests in assessing the progress of learning. He suggests that Stewart's case rests partly on a misunderstanding of a comment by Deese regarding the not (necessarily) logical nature of associative structures. As regards the nature of the links revealed by WATs, these are operationally defined by the mathmatical model (see Section 5.2.6) used to calculate relatedness coefficients from the overlap of a student's lists of responses to words representing the concepts. While the nature of the relationship between the concepts is not revealed, the presence or absence of such a link and an estimate of its strength is nevertheless of value. It may be assumed to provide an estimate of the **accessibility** of one idea from another and indeed Cochaud (1979) has used single response WATs to develop maps in which directed links are labelled "is accessible from". The accessibility of particular concepts is clearly a necessary, though not a sufficient, condition in many PS situations. Thus the findings of McCabe (1977), and of others at the University of East Anglia, that a student's PS errors are often associated with his failure to make use of ideas which other tests indicate that he knows, can be interpreted in terms of the absence or weakness of the relevant associative ("accessible from") links. It was concluded that associative mapping was likely to contribute significantly to the interpretation of the PS behaviours and that Stewart's criticisms were of limited relevance in the context of the present study.

As regards Sutton's references to the apparent randomness of some WAT responses, Schaefer has noted the same thing but apparently found enough consistency to satisfy the needs of his investigations and has continued to use the method (Schaefer 1979a, b). A few idiosyncratic associations appeared unlikely to have a major effect on inter-concept overlaps and could, in any case, be ignored in the scoring of the tests if necessary. Problems of this sort have not been reported where constrained, rather than free, associations have been employed and Sutton's comments were not felt to constitute any objection to the employment of WATs for generating individual and group maps reflecting semantic proximities.

5.5 The execution of the second phase of the study

The sections which follow deal with the construction and administration of the instruments used in the second phase of the study.

5.5.1 The construction of the PS test. The test employed in this phase of the study consisted of 30 non-mathematical items based on questions of the type used in JMB short-answer O-level chemistry papers. Four versions of the test were constructed labelled A, B, C and D. Seven "marker" items were common to all four versions, the remainder being written in different formulations and distributed amongst the test versions in a matrix design. Each test therefore contained a balance of items with reduced, standard, augmented (relevant) and augmented (irrelevant) levels of information. The structure of each test is given in Table B1 in Appendix B. Further details of the design and construction of the tests are given below.

(i) **Item selection.** In order to continue to simulate O-level examination conditions as closely as possible, and to avoid unnecessary difficulties regarding subject-matter coverage, it was planned to carry out the study with Form V students in the few weeks immediately preceding their O-levels. In view of the limited time which schools could reasonably be asked to make available with such classes, and of the need to collect other data, it was decided to restrict the test to about 30 items.

Items were selected to meet the needs of both the task-formulation study and the cognitive mapping study. Both demanded a representative coverage of tasks and, more particularly, the selection of items which could be expected to elicit a full range of PS strategies. However, while it was necessary for practical purposes to limit the range of subject matter covered by the cognitive mapping exercise (see Section 5.5.3 below), it was considered desirable to study the influence of task formulation over as wide a range of topics as practicable. A compromise was achieved by selecting three fairly broad overlapping topics (electrolysis, the periodicity of atomic structure and redox) for the mapping study. These were covered in four short-answer questions comprising 21 items. Three further questions, comprising the remaining nine items, were then added to broaden the subject matter coverage as far as possible.

A further consideration in selecting items for the test related to the intention to make comparisons; in the task formulation study between groups of students attempting different versions of the same task, and in both studies between more and less successful candidates. On both counts it was decided to include a number of "marker" items which would be common to all versions of the tests. A score based on these items would enable the equivalence of sub-populations attempting each version of the test to be established, and could be used as a co-variable in the analyses of variance if necessary (see Chapter 6). As a secondary function, the marker items would also provide an independent means of identifying strong and weak candidates, irrespective of task formation. It was clearly desirable that markers should be (a) representative of the test as a whole, and (b) sufficient in number to provide reliable data without unduly limiting the number of variable items. In practice seven marker and 23 variable items were used. The majority of items, both variable and marker, were carried forward from the first phase of the study and were supplemented by similar items following pretesting. Insofar as possible items, and particularly marker items, were selected which had high indices of discrimination. Table 5.2 shows the fairly satisfactory distribution achieved for all items as regards both the tasks demanded and the strategies which they were expected to elicit. Table B2 in Appendix B shows the facility and discrimination indices of all items at pretesting and during the study itself.

<u>Table 5.2 Distribution of items across tasks and anticipated strategies</u>

	Tasks (key - see Table 3.1)						Anticipated dominant strategy				Total
	1	2	3	4	5	6	S1	S2	S3	Mixed	
All items	1	9	9	5	3	3	11	13	3	3	30
"Variable" items	-	8	8	3	2	2	7	11	3	2	23
"Marker" items	1	1	1	2	1	1	4	2	-	1	7
"Mapping" items	1	8	5	3	3	1	8	11	2	-	21

(ii) **Task formulation.** The task formulation parameter "level of information" was varied on all the items employed in the test except those designated as markers. The four levels identified in the network analysis presented in Figure 5.1 were used in developing versions of items which were described as reduced, standard and augmented. The main characteristics of these types are shown in Figure 5.2 and are elaborated below.

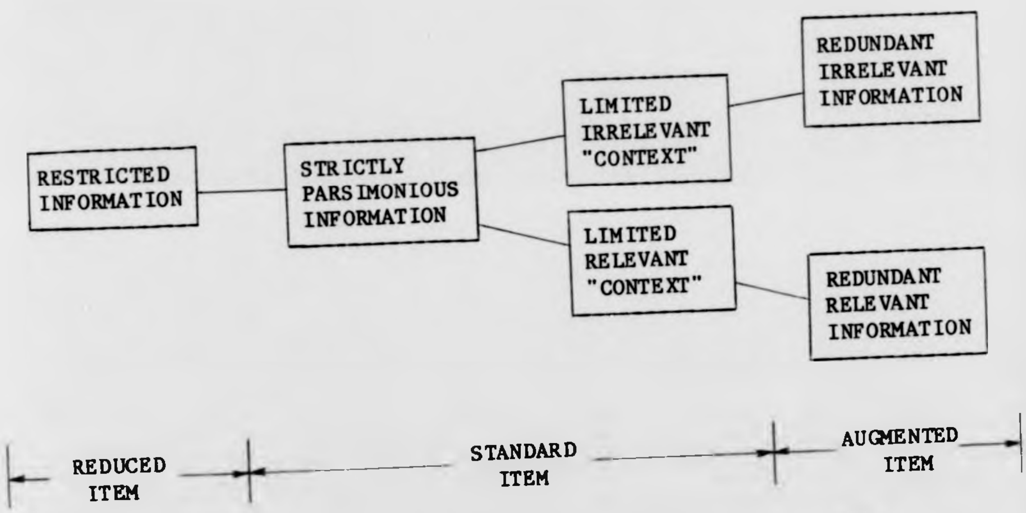


Figure 5.2 Levels of information in three types of item

Standard items, as actually encountered in O-level papers, may be strictly parsimonious or may include limited contextual information. "Parsimonious" items contain all the information necessary to define the task, including essential limitations or constraints, and no redundant information. They represent particular tasks stripped down to their essentials. "Limited context" items contain all the information of the parsimonious version and, in addition, limited information which may be loosely described as contextual. Such extra, strictly redundant information, may, for example, be involved in linking together a number of items into a several-part question on a particular theme. The redundant information may be either **relevant** (in the sense of providing potentially helpful cues, alternative solution routes, or a means of confirming a solution) or **irrelevant** (either neutral with regard to the solution, or tending to distract). Examples range from the fairly marginal case exemplified by using full equations to test a student's ability to recognise examples of oxidation involving the exchange of electrons (where a more parsimonious version of the task providing fewer potential routes to the solution is possible) through cases where a relevant or irrelevant "story-line" is used to link items representing a particular theme, to instances where examples of classes may be identified from tabulated data which includes a number of redundant cases, presumably to make guessing less profitable.

Augmented items differ in degree rather than in kind from the "limited context" items described above. They are items of a kind not normally encountered in examinations and were compiled by adding redundant information to a standard item and going beyond the "limited context" described and exemplified above. Two sub-classes of such items were written using added information which was relevant (as defined above) or irrelevant (also defined above).

Reduced items differed from the strictly parsimonious type described above in that essential information was withheld. Such items varied from cases in which the missing information could reasonably be assumed (eg in asking the name of an industrial process using a particular raw material and a given chemical process, but omitting the name of the product which might or might not be regarded as strictly essential), to cases in which an essential limitation or condition was omitted but in which students might be able to make assumptions of a neutral kind (eg in failing to state the nature of the electrodes in a task relating to the electrolysis of an aqueous solution). Although Krutetskii (1976) employed items in which there was insufficient information for any rational attempt to be made in an individual interview situation, the use of such extreme items in the context of a written test was judged to be undesirable in view of the strong risk that students might be unduly confused to the detriment of their performance on the test as a whole.

The four item-types may be further exemplified by reference to an example from the test in Table 5.3. This item demands only the simple recall of the name of an industrial process. The augmented (relevant) version adds further information of a kind which may help to cue the answer, while the information added in the augmented (irrelevant) version is both redundant and neutral as regards any likelihood of cuing. The reduced version is potentially ambiguous, although in the context of the syllabus only one answer is likely; the absence of the word ammonia probably removes the strongest cue to the answer.

<u>Table 5.3 Four versions of item 3(c)(ii)</u>

3(c) Give the names of the industrial processes which involve the following .... (other parts omitted) -

<u>Standard:</u>	(ii) the catalytic reduction of nitrogen to ammonia.
<u>Augmented (relevant):</u>	(ii) the catalytic reduction of nitrogen to ammonia using a catalyst of finely divided iron with promotors.
<u>Augmented (irrelevant):</u>	(ii) the catalytic reduction of nitrogen to ammonia in the gaseous phase.
<u>Reduced:</u>	(ii) the catalytic reduction of nitrogen.

(iii) **The compilation of the test.** There is little to be added regarding the compilation of the test. Although the example given in Table 5.3 is straightforward enough, limits to the task formulation study were set by intuitions regarding the range of formulations which were credible for a given task and by the author's ingenuity in devising such formulations. Thus, although 22 of the 23 non-marker items were produced in satisfactory augmented (relevant), as well as standard, versions, only 15 augmented (irrelevant) and 12 reduced items remained after all the versions of all the items had been reviewed critically with a chief examiner in chemistry.

In compiling the four versions of the test necessary to cover all the formulations of certain items, it was considered important to avoid tests which would differ systematically. Thus, for example, if pupils had encountered a succession of items containing insufficient information, a cumulative effect, perhaps of confusion or exasperation, might well have introduced an unwanted factor into the investigation. The test versions were thus put together using a matrix design so that, not only were the same basic tasks encountered in the same order in every test, but so that each contained, as far as possible, a similar distribution of standard, reduced, augmented (relevant) and augmented (irrelevant) versions. Where an item had three versions only, the standard version appeared in two of the tests (except in one case, where the interaction of a particular item with a stem shared by other items, made it expedient to duplicate the reduced version). In the only case where an item had only two formulations, each appeared in two versions of the test. The distribution of task formulations amongst the test versions A, B, C and D is given in detail and in summary, respectively, in Table B1 (Appendix B) and in Table 5.4 (below). Copies of the four versions of the test, which were printed on different coloured paper for ease of recognition, will also be found in Appendix B.

Table 5.4 Distribution of different task formulations across four test versions

Task formulation; (text code)	Test version; text code (computer code)			
	A(1)	B(2)	C(3)	D(4)
Standard, (S)	9	10	10	11
Reduced, (R)	3	3	4	4
Augmented (relevant), A(R)	7	6	6	4
Augmented (irrelevant), A(I)	4	4	3	4

5.5.2 **A student's self-reporting form.** Mention must be made of a largely unsuccessful attempt to obtain direct information concerning students' strategies through a self-reporting form. The intention had been that students, on completing each item in the test, should record the strategy they had employed by ticking the appropriate "box" on a specially designed form. A simplified set of strategies, based on the findings of the first phase of the study, was used for this purpose. Students were briefed on these before starting the test. They were also asked to show any working involved in answering the item, make a brief note of any difficulty encountered and to indicate their confidence in their answer using a simple four-point scale. It had been hoped to validate the self-reporting form during preliminary trials in a number of ways:

(i) by reference to the strategies anticipated for each item on the basis of the first phase of the study;

(ii) by comparing a student's assessment of his own strategy with evidence of working from his written answer;

(iii) by interviews with individual students immediately after they had used the forms, and

(iv) possibly by asking a small number of students to make self-recorded PS protocols while attempting the items and completing the form.

A start was made while pilot-testing the additional items required for the second phase of the study. The students involved appeared to enjoy using the form and, during a short class discussion afterwards, claimed that it required little effort and did not detract from answering the items. This appeared to be confirmed when their results were found to match the predictions of their chemistry teacher very well, both as regards levels and rank orders of performance. However, the students' assessments of their own strategies did not match well with the author's expectations and indeed many were not feasible. Little evidence was available from written work and plans to interview individuals after the test had fallen through. However, it seemed clear that the students had not understood the categories of strategy described, and a slightly simpler scheme was devised with clearer instructions. Unfortunately it proved impracticable to arrange further trials before the main study was carried out. The original pilot work had suggested, however, that use of the form would not interfere with students' performance. The self-reporting form was therefore used during the main study in the hope that validation might subsequently be established in separate trials. In the event, inspection of students' responses soon showed that the original difficulties with the instrument had not been overcome by the changes made and further work had to be abandoned. Nevertheless, useful information was available from some students on some items in the form of the brief notes they were asked to make about their working and their difficulties. While this was too patchy to make systematic analysis worthwhile, it was hoped that such information might contribute to the interpretation of particular PS episodes.

Copies of the instructions to students, of a summary key for the self-reporting form and of the report form itself (which was printed in the test booklet on pages facing each item) are included in Appendix B.

5.5.3 **The construction of the word association test.**

(i) **The selection of stimulus words.** It has been asserted (Deese, 1965) that meaningful associative structures can be obtained using stimulus words selected "by almost any method that takes advantage of a relationship among concepts and words in the language". Deese (1962), for example, used words which all occurred as frequent responses to a given stimulus word. Johnson (1965) used the frequency count of words representing concepts in Newtonian mechanics in a student's text and Shavelson (1974) used "key concepts" selected as critical to the subject-matter structure of a topic in mathematics. As the aim of the current study was to examine relationships between students' associative structures and their performance on particular problems, stimuli were selected with reference to problem-relevant chemical concepts. The first four questions (21 items) on the PS test were contrived to provide an adequate sampling of PS behaviour while reflecting a relatively limited set of overlapping topics; namely electrolysis, the periodicity of atomic structure, and redox. The use of four questions also ensured, through the matrix design of the tests, that the different task formulations would be adequately sampled. The text of these questions was found to include the names of 21 problem-significant chemical concepts, and to these were added the names of six further concepts which the first phase of the study indicated were likely to be involved in solving the items. The final list appears in Table 5.5.

While the quantitative mechanics concepts employed by P E Johnson, by Shavelson and by Preece (among others) in similar studies, are tightly inter-related by defining equations and can thus be represented by relatively simple logical structures (eg Preece, 1976b), such was not the case for the sets of chemical concepts involved here, which tended to be inter-dependent in more complex ways. Less highly organised structures were, however, put together for each of the three topic areas selected. Each concept was linked to those other concepts which the author would have expected to be used in defining or describing it within the context of the O-level curriculum and of the problems. The resulting structures appear in Figures C1 to C3 in Appendix C. Although these structures cannot claim to be unique representations of each topic area, they provided a useful a priori basis for planning the tests and for the qualitative evaluation of students' cognitive maps.

Table 5.5 Ordered list of stimulus words employed in the WAT

1. Electrolysis
2. Halogen
3. Reduction
4. Isotope
5. Negative
6. Formula
7. Anode
8. Oxidation
9. Periodic table
10. Electron
11. Electrolyte
12. Positive
13. Atomic mass
14. Electrode
15. Proton
16. Electrochemical series
17. Group
18. Ion
19. Atomic number
20. Charge
21. Element
22. Electronic configuration
23. Reducing agent
24. Valency
25. Cathode
26. Electric current
27. Atom

A preliminary trial was held using three of the four questions to be studied with a reduced WAT of 22 words. A simple hierarchical analysis produced an obviously meaningful associative structure for the group. The dendogram is included as Figure C4 in Appendix C . Although no fuller analysis was attempted owing to the preliminary nature of the work, this trial provided adequate encouragement to continue with the approach described.

(ii) The order of the stimulus words. Although Deese (1965) regards WATs as being close to context-free, it is regarded as desirable to try to minimise "priming". This is an effect whereby a word recently encountered as either a stimulus or a response, may influence a subject's associations to a new stimulus word. This effect was eliminated in a few early studies by alternating key concept words with irrelevant words from a different topic area (eg Johnson, 1964). In the present case this would have resulted in an unacceptably lengthy list, and such precautions have not been considered necessary in most recent studies. It has been reported that the influence of this sort of priming can be eliminated by as little as 15 seconds of interpolated meaningful verbal activity (Cramer, 1966) and that priming does not occur for associative habits of zero strength (Cramer, 1968). With these findings in mind, the order in which stimulus words appeared in the test, while otherwise random (a random number generator was employed), was constrained so that no two concept words which would normally be used in defining or describing one another (ie words directly linked to one another in the content structures presented in Appendix C) should appear in adjacent positions. This was deemed likely to minimise any possible stimulus to stimulus priming and parallels the procedure employed in a recent study by Preece (1978a). The use of different random orders on different copies of the test has been favoured in studies concerned with group characteristics. However, in view of the potential interest of individual differences in cognitive structure, the same random order was used in all the lists. Preece's (1978a) study again provides a precedent.

(iii) **The time allowed for responding.** It was decided to allow students 30 seconds for responding to each stimulus word and this requires explanation. Most reported work with both free and constrained associations has employed 60 seconds following the pioneering study of

Noble (1952) who quoted the judgement of his subjects during pre-testing in justification and claimed that one minute allowed "a reliable time sample of the subject's response hierarchy". However, although using young adults as subjects, Noble found it desirable to employ rest periods after every 36 words and Preece (1976a) refers to problems of "decreasing motivation and fatigue" in a study of only 15 words which he administered to students ranging from about 12 to 23 years of age; he also mentions the need to caution students against frivolous responses. Thus, although a 27 item WAT with a 60 second responding time could have been administered in the single school period available for this part of the work, it was felt that a shorter time interval might be desirable. Laffal and Feldman (1962) allowed only 25 seconds for (oral) responses to a WAT (and employed rest periods every 12 words) and, while they do not justify this time interval, obtained satisfactory results using it. It is relevant that P E Johnson (1967) found that "constrained" responses (ie responses appearing with the corresponding stimulus word in a relationship defining either, or together defining a third stimulus word) occurred significantly more frequently in the top half of the response hierarchy of high, as opposed to low, achievers. Shavelson (1973) found similarly that the number of "constrained" responses in the upper half of students' lists was the best predictor of performance on a relevant post-test, and that the numbers of such responses in the lower half of the list was a negative indicator of achievement. It therefore seemed probable that, in PS situations, a student's more spontaneous and immediate associations were likely to be of greatest importance.

With this in mind, and because of concern about the possible influence of fatigue effects, students were closely observed during the pilot trials referred to above. A 60 second responding time was used but students were asked to mark each of their response lists when a time-check was given after 30 seconds. A fatigue factor was readily, if subjectively, apparent, as a degree of restlessness in the class towards the end of each 60 seconds period; increasingly so towards the end of the test. This was strongly confirmed in an evaluative discussion with the class immediately afterwards. Many claimed to find the test long and boring (22 words only were involved) and virtually all felt that 30 seconds per word allowed ample time for what they regarded as important responses. Subsequent analysis showed that (a) on average nearly 77% of responses were made during the first 30 seconds, and (b) that the responses made after 30 seconds were much more likely to be "unsatisfactory" than those made earlier. In this context, unsatisfactory responses included vague, poorly discriminated associates (the words "chemistry" and "reaction" tended to occur towards the ends of some lists, apparently almost irrespective of the stimulus word) and failures of set ("chaining" and the use of conscious strategies in generating further responses). It seemed clear that many students tended to run out of spontaneous associations in less than 30 seconds and that some tended to react to this in ways which were undesirable from the point of view of the investigation. It was therefore decided to limit the response time to 30 seconds in the main study.

(iv) **The format of the WAT.** The format of the WAT was closely modelled on those described in turn by P E Johnson, Shavelson and Preece. A booklet was produced in which a page of instructions and a completed example were followed by 27 pages, each headed by a different stimulus word. The word was repeated 12 times down the left-hand side of the page to encourage continued association rather than chaining. A context was provided by the title "Word Association Test - Chemistry" on the cover and by the fact that the test was carried out during a chemistry class. However, there was no formal constraint limiting responses. This follows Preece's (1976a) report that such formal constraints might tend to mask individual differences. A copy of the test, including instructions, will be found in Appendix C.

5.5.4 **The administration of the instruments.**

(i) **The test population.** The tests were designed to be used with JMB O-level chemistry candidates who had completed the chemistry syllabus. Because of the novel instructions involved in using the students' self-reporting form (although in the event this generated only limited useful data), it was necessary for the test to be administered personally by the author. This limited the number of schools approached to ten, of whom seven participated during the second half of the spring term 1980, some two to three months before the final O-level examinations.

In planning the administration of the test, it was necessary to consider the nature of the intended analysis. The cognitive mapping exercise did not impose any serious constraints in this respect. However, in studying the influence of task-formulation, a series of analyses of variance were to be carried out. In planning such analyses in educational research involving more than one school, it is normal to include "school" as a factor on the grounds that any particular performance is likely to vary betwen schools (Kerlinger, 1964). Failure to do so may result in any systematic variance associated with a particular treatment (or in this case, task formulation), being "swamped" by that between schools and thus not being recognised as significant. Although there is no absolute constraint on cell size in analysis-of-variance designs, Kerlinger suggests that it is usual to aim for a minimum of about 10 subjects. Given four test versions, this implied a group of 40 or more students at each school. Two of the schools were able to provide about 40 students each and one provided 80. However, the remaining four schools were entering smaller numbers of candidates. It was therefore decided, for statistical purposes in the analysis of variance only, to treat these latter schools as a single group. Unfortunately the group was heterogeneous and representative rather than homogeneous and is difficult to justify as an extra "school" except in terms of increasing the total numbers in the analysis without interfering with the recognisable groups (schools) of adequate size. Since about 80 students were involved this was considered adequate justification.

The focus of interest in the analyses was to be with main effects attributable to "level of information" and with interactions involving high and low performers. There was no strong rationale for considering gender as a significant factor in the design, nor would it have been practicable to do so with the population available. However there had been weak indications, in the first phase of the study, that the PS behaviour of girls might differ from that of boys in certain respects. Information regarding gender was therefore sought as supplementary data.

Each participating school was assigned a reference letter and the distribution of students by test version, by school and by gender, after losses due to experimental mortality, is shown in Table B3 in Appendix B.

(ii) **The administration of the WAT.** Johnson (1965) reported that the administration of a WAT following a PS test led to a significant increase in problem-relevant associations. The WAT was therefore administered from one to five days **before** the PS test (the longer interval being due to weekend and/or half-term holidays). It was assumed that a day or more's interpolated activity would be sufficient to render any facilitating effect on the subsequent PS test (Johnson, 1965) negligible. Testing was programmed so that no chemistry lessons intervened between the two tests. It thus seems reasonable to assume that students' relevant associative structures remained fairly constant over the interval involved, and that the cognitive structure mapped by the WAT was essentially the same as that available to the student when subsequently attempting the PS test.

The WAT was administered by the author following procedures similar to those described, for example, by Shavelson (1972) and by Preece (1976a). After students had read the instructions, they were asked to turn over the page and to look at a completed example involving a chemistry word which was unrelated to the concepts in the test. This was discussed briefly, emphasising the need to repeat the "key word" to themselves before every response. They were asked to avoid (a) repeating responses, (b) "chaining", (c) the use of explanatory phrases, and (d) the use of deliberate strategies to generate responses. Each of these was briefly elaborated and the importance of spontaneity in responding was emphasised. These points had been compiled as a set of notes entitled "Word Associaton Test - Chemistry: Procedure for Test Administration" of which a copy will be found in Appendix C. They were closely paraphrased rather than actually read out to each class as this enabled a better rapport to be established with the students. The students were encouraged to ask questions but in general little clarification was needed. Once the instructions had been discussed (about five minutes was usual) students were told they would be allowed 30 seconds for each word and that, although 12 spaces would be provided on each page, this was neither an upper nor a lower limit, although they should work as fast as possible. They were then told to turn over the page and start working, the former instruction being repeated at 30 second intervals. An interval of about five seconds was allowed for students to finish writing and turn over each time.

At the time of administering the WAT, the chemistry teacher was asked for the students' mock O-level marks to provide additional baseline data.

(iii) **The adminsitration of the PS test.** The test was administered in every case by the author. For each school, or group of schools, equal numbers of the four versions of the test were arranged in random order using a random number generator. All the students participating from any particular school took the test at the same time. When they were seated, the tests were distributed from the top of the randomised pile, up and down alternate rows, from right to left. They were instructed to read the front cover and to fill in their name and gender in the space provided.

Students were told that the test would be marked and that the marks would be sent back to the school. However, they were reassured concerning their anonimity in regard to research reporting. They were then asked to turn over the cover and read the instructions relating to the self-reporting form. The author spent about 15 minutes going through these instructions and answering any questions raised by the students concerning them. As with the WAT, a set of notes entitled "Procedure for Test Administration" had been compiled to ensure uniformity in the administration procedure. As before, these were closely paraphrased, rather than being read verbatim, so as to establish a better rapport with the group. The test was then completed under normal examination conditions. A generous time allowance of 60 minutes ensured that virtually all students completed the test without undue strain. Many finished in 45 minutes, most in 50 minutes and all but one in 60 minutes.

CHAPTER 6

THE INFLUENCE OF "INFORMATION LEVEL" ON PROBLEM-SOLVING PERFORMANCE

The influence of the level of information provided in an item on PS performance was examined by analysis of the variance of students' scores on the different versions of the items described in Chapter 5. Before proceeding to an account of this in Sections 6.2 to 6.4, the conduct of a number of preliminary analyses of data relating to the tests is described in Section 6.1

6.1 Some characteristics of the PS tests and of the group of students involved in the study

Following the marking of the test and an examination of the statistical characteristics of, in particular, the "marker" items common to all four versions of the test, students' scores on the latter were used to assess the effectiveness of the population sampling (that is, the equivalence of the groups of students attempting each version of the test). The effectiveness of the matrix sampling of the different versions of various items (that is, the equivalence of the four versions of the test from the point of view of overall difficulty) was also examined. Finally, evidence was considered regarding the reliability and validity of the tests. A detailed account of these preliminaries follows.

6.1.1 **The marking of the tests and the performance of the "marker" items.** Although the test was not of the "objective" type, marking presented few problems and only minor adjustments had to be made to the prepared marking scheme. The final version of this, together with comments on the items for which changes were made, are given in Appendix B.

Following marking, the students' scores on each item were transcribed onto coding sheets and punched onto computer cards. Each student was assigned a reference number, and school, class, gender, test version (A, B, C or D) and mock O-level scores (MOCK) were also encoded. Total test scores (TOTAL) and separate sub-scores on the sets of marker items (MARKRTOT) and variable items (VARTOT) were computed as additional variables prior to analysis. All analyses were carried out using appropriate programmes from the Statistical Package for the Social Sciences (SPSS), versions 6.5 and 7.0.

An item analysis was carried out to provide general information regarding the performance of both variable and marker items. Since the variable items were to be used to explore the influence of task-formulation on PS behaviour, there was no call on the usual norm-referenced function of this analysis in eliminating unsuitable items. However, some preliminary indication of items behaving in an unusual manner was considered relevant and it was essential to confirm that the marker items performed satisfactorily.

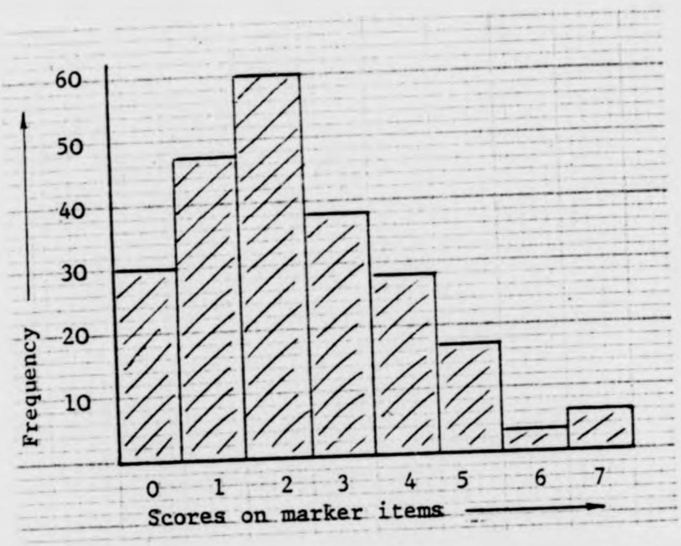
The facilities (p) and indices of discrimination (d) of the standard versions of each item are given in Table B2 in Appendix B. Also shown in this table are the "p" and "d" values for these items derived during the first phase of the study and/or during pre-testing. The methods used in calculating these indices were described in Chapter 4.

In general, facility levels were somewhat lower than had been anticipated from earlier experience with the items; nevertheless in most cases the value of "p" fell into the normally accepted range. Discrimination was generally satisfactory with the value of "d" on almost a third of the items equalling or exceeding 0.40. Together with the overall distribution of scores (described in Section 6.1.4) this suggested that the test could be expected to discriminate adequately between the more and less successful problem-solver. The seven marker items discriminated particularly well (the average value for "d" was 0.48), confirming their suitability both as a means of distinguishing successful from unsuccessful problem-solvers independently of the variable items, and as the basis of an independent variable in an analysis of co-variance if necessary.

Four of the variable items (4(c)(ii) and (iii), 5(c) and 7(c)) would have been judged unsatisfactory in a norm-referenced test as regards both facility and discrimination. Discussion of all of these except 5(c) will be found in the notes concerning cases of difficulty in marking in Appendix B; it should be noted that item 5(c) had performed well in the first phase of the study (p = 0.35; d = 0.85) and no obvious defects were apparent on inspecting the text of the item. Two other items (1(c) and 4(a)) discussed in Appendix B were rather difficult but discriminated well, while a third (5(b)) was of average facility but discriminated poorly. Some reservations regarding the wording of 1(c) are also recorded in Appendix B; however all but two of the items mentioned had been pre-tested satisfactorily and none appeared to be unsound.

Before proceeding with the analysis, it was necessary to establish the reliability and validity of a student's total score on the sub-test of seven marker items (MARKRTOT) which were common to all four versions of the test. Distribution of the scores is shown in Figure 6.1, together with a tabulation of the mean, standard deviation and reliability (Kuder-Richardson formula 20) of this sub-test.

Despite the positive skew, the distribution of scores was regarded as satisfactory and it was clear that the test discriminated quite effectively, especially amongst the higher achievers. The standard deviation was between a quarter and a fifth of the range, which is normal, and the reliability of 0.60 was considered acceptable for a short test.



Mean (ex 7)	Standard deviation	KR 20 Reliability	N
2.35	1.68	0.60	229

Figure 6.1 Distribution and statistics relating to total scores on marker items common to all versions of the PS test

The test had good face validity as a measure of the type of performance being investigated, covering a representative range of tasks and expected strategies (see Table 5.2, p 194). Its concurrent validity was assessed by correlation with (i) the mock O-level marks awarded by the schools (MOCK) and (ii) overall performance on the variable parts of each test (VARTOT). Because each school had set and marked its own mock examination, product moment correlations involving MOCK were calculated separately for each school and then averaged using Fisher's "z" transformation, as described by Lewis (1967).[1] Similarly the equivalence of the four versions of the test had yet to be established, so correlations involving VARTOT were calculated separately for each version and then averaged.

1. Product moment correlations were preferred to rank order correlations because no equivalent averaging procedure is available when the latter are used.

Tables 6.1 and 6.2 show that both the overall correlations and all the individual coefficients, were positive and highly significant statistically, although the relatively low r-value associated with test version A was noted. This suggested that a student's score on the marker items might be regarded as providing an acceptable estimate of his PS performance in the context both of O-level chemistry examinations, and of the particular instruments employed in this study.

Table 6.1 Product moment correlations of MARKRTOT with MOCK

School	C	E	K	H	L	G	J	Mean
Correlation	.58*	.55*	.46*	.45+	.76*	.67*	.61*	.59*
N	76	35	39	13	25	18	23	

Table 6.2 Product moment correlations of MARKRTOT with VARTOT

Test version	A	B	C	D	Mean
Correlation	.28+	.45*	.57*	.74*	.53*
N	56	58	60	55	

Key (Tables 6.1 and 6.2): N number of students
\+ significance level $p < 0.05$
* significance level $p < 0.01$

6.1.2 **Some characteristics of the group of students participating in the study.** The assignment of students at random to one of the four versions of the test in each school, or group of schools, was described in Chapter 5. Out of 242 students involved, 229 took both the WAT and the PS test. The results of the remainder were discarded to maintain the integrity of the link between the PS test and the cognitive mapping study. This experimental mortality left the final distribution of students, as shown in Table B3 in Appendix B, a little uneven. In particular, the proportion of girls to boys attempting version A of the test appeared rather high. However, chi-squared tests confirmed that the distribution of test versions by school or group, and by gender, did not differ significantly from random.[1]

In order to facilitate the planning of the subsequent analyses of variance, the variability of the students' PS performance by school or group and by gender were investigated. Tables 6.3 and 6.4 show the breakdown of performance on the marker items by, respectively, school or group, and gender. From the results it is clear that the school or group was a significant source of variance. In view of the fact that the

Table 6.3 Breakdown of MARKRTOT by school or group

School or group	C	E	K	X	
Mean score (ex 7)	2.84	2.91	1.51	2.04	F = 8.42
Standard deviation	1.62	1.80	1.12	1.68	df = 3
N	76	35	39	79	p < 0.001

Table 6.4 Breakdown of MARKRTOT by gender (mixed schools only)

Gender	Female	Male	
Mean score (ex 7)	2.55	2.58	F = 0.02
Standard deviation	1.59	1.79	df = 1
N	94	96	p > 0.10

low-performing school K was an all-girls school, this group was omitted from the breakdown by gender to avoid confounding effects due to gender and to school. The results showed, that within mixed schools or groups, gender was not a significant source of variance.

1. For test version by school or group, $\chi^2 = 0.94$, df = 9, $p > 0.10$; for test version by gender $\chi^2 = 4.31$, df = 3, $p > 0.10$.

For purposes of subsequent analysis, it was also necessary to investigate the effectiveness of the population sampling with regard to the four versions of the test. Table 6.5 shows the breakdown of students' performance on the marker items by test version. The mean

Table 6.5 Breakdown of MARKRTOT by test version

PS test version	A	B	C	D	
Mean score (ex 7)	1.88	2.62	2.30	2.60	F = 2.48
Standard deviation	1.14	1.85	1.70	1.83	df = 3
N	56	58	60	55	p = 0.06

score achieved by students to whom test version A had been assigned was noticably lower than those achieved by students who had received the other tests. Although this result marginally failed to meet the accepted test of statisticial significance at the 5% level, the remaining analyses were carried out on the assumption that, despite random sampling, the students who received test version A might be slightly less effective problem-solvers than the others involved in the study.

6.1.3 **Some characteristics of the different test versions.** The matrix sampling of the different task formulations in the four versions of the test was described in the last chapter. This was undertaken to ensure that each version represented a similar challenge to the students to whom it was assigned, thus avoiding systematic influences due to differences in the overall properties of the different versions. Because the apparent non-equivalence of students' abilities across the four versions (reflected by the lower performance of the students attempting version A on the marker items; cf Table 6.5) failed to reach accepted levels of statistical significance, total scores on the test were simply broken down by test version in the first instance. Table 6.6 shows that the test versions appeared to differ significantly in difficulty. However, in view of the probability (0.94) that the students

Table 6.6 Breakdown of total score by test version

PS test version	A	B	C	D	
Mean score (ex 30) Standard deviation N	9.36 3.68 56	12.41 4.39 58	11.40 4.82 60	12.80 4.90 55	F = 6.63 df = 3 p < 0.001

receiving version A were less effective problem-solvers than the remainder, and of the rather high proportion of girls on version A (0.69 compared to 0.58 overall, cf footnote p 215 and Table B3, Appendix B), a more rigorous test was considered necessary. Total scores on the variable items (VARTOT) were subjected to a three-way analysis of variance with respect to test version, school or group, and gender, using MARKRTOT as an independent co-variable; the reason for the latter was to extract variance due to the different levels of student ability as reflected by scores on the marker items. To avoid uninterpretable effects due to the all-girls, low performance school K, this group was omitted. The full results of the analysis are shown in Table B4 in Appendix B. There were no significant two or three-way interactions and no main effect due to school or group. A main effect due to gender, significant at the 5% level, showed that girls (from the mixed schools) averaged 0.93 marks less than boys on the 23 variable items. Of more interest was the main effect due to version which is summarised in Table 6.7. This shows that there was an overall effect due to test version, version A being rather more difficult than the others. This was unfortunate insofar as statistics using total test scores irrespective of version could not strictly be justified (but see Section 6.1.4 below). However,

Table 6.7 Summary of main effects on VARTOT attributable to test version following three-way analysis of co-variance

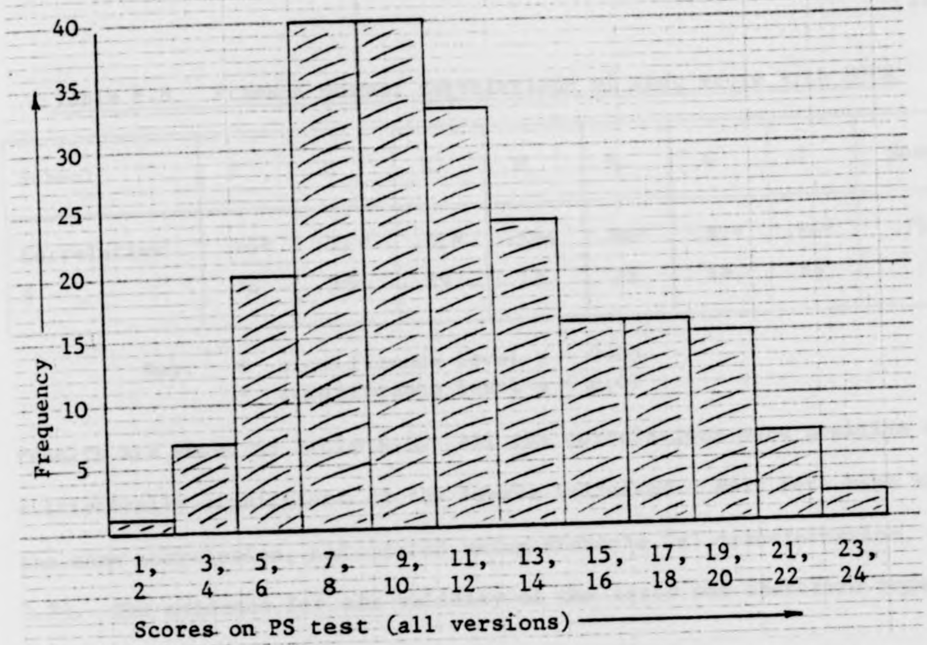
PS test version	A	B	C	D	F = 2.83
Adjusted mean score	8.54	10.01	9.83	10.24	df = 3
N	46	48	50	46	p < 0.05

the absence of interactions was encouraging and showed that the variable parts of the test, taken as a whole, did not lead to different patterns of performance by the students from different schools and/or of different gender. The small difference in overall difficulty between the test versions, in the absence of such interactions, did not in any way invalidate the item by item analysis of variance subsequently undertaken.

6.1.4 **The reliability and validity of the PS test.** Because the four versions of the test were not of exactly equivalent difficulty, the distributions of marks and reliabilities of each had to be considered separately. Similarly, the assessment of concurrent validity by correlation with mock O-level scores had to be undertaken separately for each version. However, for ease of reporting, the results calculated across tests will be presented here. The approximation involved in doing this is slight and the results do not differ materially from those obtained by more rigorous procedures.

The distribution of total scores is shown in Figure 6.2 together with a breakdown of the mean, standard deviation and reliability (Kuder-Richardson formula 20) of the test. Despite the positive skew, the distribution was regarded as satisfactory and the test clearly discriminated quite well, particularly amongst the higher achievers.

Distributions for test A were somewhat more, and for the other tests, somewhat less, positively skewed than that shown in Figure 6.2. The



Mean (ex 30)	Standard deviation	KR 20 Reliability	N
11.49	4.64	0.76	229

Figure 6.2 Distribution and statistics relating to scores on the PS test (all versions)

standard deviation was slightly less than a fifth of the range of scores which is normal for a population of this size (Ebel, 1965). The Kuder-Richardson reliability of 0.76 is also satisfactory for a test of 30 items (those on the separate versions were 0.63 for version A, 0.78 for version D and 0.79 for both versions B and C).

The test had good face validity (at least insofar as the standard versions of the items were concerned), as a measure of PS performance in O-level chemistry examinations, covering a representative range of tasks and expected strategies (see Table 5.2, p 194). Its concurrent validity was assessed by correlation with mock O-level marks supplied by the schools. Each school set and marked its own mock examination and separate correlations were therefore calculated for each school. The

Table 6.8 Product moment correlations of test score with MOCK

School	C	E	K	H	L	G	J	Mean
Correlation	.74*	.71*	.52*	.53+	.84*	.82*	.86*	.73*
N	76	35	39	13	25	18	23	

Key: + significance level $p < 0.05$
* significance level $p < 0.01$

results are shown in Table 6.8. All the correlations were positive and statistically significant at the levels indicated. Most were high and the mean correlation, calculated using Fisher's "z" transformation, was 0.73. The evidence for the validity of the tests was therefore regarded as adequate.

6.2 A general account of the analyses of variance

The purpose of the analyses was to establish whether the nature of the task formulation, specifically the level of information provided, had a significant influence on students' PS success. This was investigated with reference to scores on individual items. (In designing the tests, each particular information level was represented by modified versions of a range of items. The way in which this could be done was constrained by the nature of each item and by the need for plausibility. Thus the way in which each information level could be realised differed somewhat from item to item, and only a minority of items could be written in all four versions. In the absence of any prior evidence regarding the equivalence of the effects of the different realisations of a particular information level across different items, there would have been no justification in attempting comparisons using total scores derived from groups of items nor, indeed, was the matrix sampling design of the tests planned with this in mind.)

The analysis of students' scores on the marker items was described in the previous section. It was concluded from this that the sub-populations assigned to the four versions of the test were probably not fully equivalent in terms of PS ability. It was also established (using analysis of co-variance with scores on the marker items as an independent co-variable to allow for probable differences in ability), that the tests were not fully equivalent, version A being somewhat more difficult. However, the absence of any two- or three-way interactions in the three-way analysis of variance of total scores on the variable items, did allow the approximate assumption that such interactions would also be absent in the case of individual items. This was checked, and broadly confirmed, by repeating the three-way analysis (with MARKRTOT as co-variable) with the individual items. Table 6.9 reports the results of this exercise in qualitative terms, listing only those items for which interaction effects were significant at the 5% level or better. For only five out of the 23 variable items were any such interactions found.

Table 6.9 Items showing significant interaction effects ($p < 0.05$) on a three-way analysis of variance

Interaction effect	Items
2 way interaction VERSION/SCHOOL	502
2 way interaction VERSION/GENDER	305
2 way interaction SCHOOL/GENDER	305, 404, 702, 703
3 way interaction VERSION/SCHOOL/GENDER	none

All of these were first order (two-way) and only two involved task formulation (version).[1] These two interactions will be discussed later in Section 6.3.3.

1. Four items showed significant interactions involving school and gender. All these appeared to be associated with the relatively lower performance of girls at the all-girls school K relative to girls from mixed groups.

6.2.1 Analysis of variance of item scores by information level. The absence of second-order interactions, and the very limited number of first-order interactions involving task formulation, have been reported above. A simple one-way analysis of variance was therefore carried out in the first instance. To correct for the non-equivalance of the sub-populations, the variance of the scores on each item was broken down by task formulation (information level) with MARKRTOT as an independent co-variable. The full results of this analysis are reported in Tables B5 (1 to 23) in Appendix B. A summary is presented in Table 6.11 (p 226) in the form of a listing of the item facilities associated with each sub-version of the task, adjusted for the non-equivalence of the sub-populations (through the co-variable MARKRTOT). Supplementary two-way analyses were conducted for the two items which had shown significant first-order interactions in order to take account of the interaction variance and obtain a better estimate of the appropriate F-ratio. The variance of the scores was analysed for item 305 by task formulation by gender, and for item 502 by task formulation by school (in both cases with MARKRTOT as an independent co-variable). The results of the supplementary analyses are reported in Tables B5 (24 and 25) in Appendix B and are summarised in Table 6.12 (p 227). The results of all the analyses described in this paragraph are fully discussed in Section 6.3.

6.2.2 Two-way analysis of variance of items scores by information level by achievement level. The possibility has already been mentioned that the influence of task formulation (information level) might differ for students of high and low achievement in chemistry; in particular that the additional cuing provided in the augmented (relevant) items and the distraction provided in the augmented (irrelevant) items might, respectively, raise and depress the performance of low achievers while having relatively little effect on that of high achievers. To investigate this possibilty, two-way analyses of the variance of item scores by task formulation (information level) by achievement tertile, were carried out with MARKRTOT as co-variable.

Students were assigned to achievement tertiles according to performance on marker items. This was supplemented by reference to mock O-level scores where further discrimination was needed between students of equal rank on MARKRTOT. The composition of the sub-populations in each tertile by school, gender and test version is shown in Table 6.10.

Table 6.10 Composition of sub-populations comprising tertiles by school, gender and test version

Tertile	School/Group				Gender		Test version			
	C	E	K	X	F	M	A	B	C	D
Top	34	15	8	20	41	36	12	22	22	21
Middle	28	10	9	28	43	32	23	18	17	17
Bottom	14	10	22	31	49	28	21	18	21	17

Although there was inevitably some imbalance, the proposed analysis of variance was justified by the lack of interactions (already reported) involving task formulation with school and/or gender. The relatively small number of students from the top tertile attempting test version A militated against the employment of quartiles rather than tertiles. The former had initially been proposed on the grounds of maximising discrimination between the highest and lowest achievement groups. However, the cell corresponding to the top quartile/test version A was found to contain only five students compared to an average of between 14 and 15 in the other cells. This was not regarded as an adequate sub-population to represent performance on those item versions appearing in test A, and the analysis by quartile was therefore abandoned (although in practice yielding results closely similar to those obtained using tertiles).

The results of the two-way analyses of variance are reported in full in Tables B6 (1 to 23) in Appendix B. A summary is provided in Figure 6.4 (pp 238 to 240) and the results will be discussed in Section 6.3.2

6.3 The results of the analysis of variance

The aim of this part of the study was to examine the way in which students' PS performance may be influenced by the level of information provided in the PS task. It was hypothesised that increasing the level of **relevant** information would facilitate PS along the series represented by the "reduced", "standard" and "augmented (relevant)" versions of items; conversely that increasing the level of **irrelevant** information from the "standard" to the "augmented (irrelevant)" format, would depress performance. It was also hypothesised that these effects would be less marked for students of high achievement. The results of the analyses of variance described in the last section are reported below in this context.

To complement the statistical results, reference will be made to qualitative data obtained from students' scripts and self-reporting forms. For this purpose 10 scripts were selected at random (using a random number generator) from each performance tertile on each version of the test, 120 scripts in all. The amount and quality of information available from this source was generally poor, though varying widely from student to student and from item to item. For this reason no data of a more quantitative sort can be presented. However, the information relating to particular cases was useful in the initial interpretation of findings relating to particular items and will be reported anecdotally where relevant.

6.3.1 The influence of information level on performance. The results of the one-way analysis of variance of item scores (facilities) by information level with MARKRTOT as co-variable are presented in Table 6.11. Supplementary data from relevant two-way analyses, for two items which showed interactive effects involving information level, are shown in Table 6.12.

When examining the results, very small differences of facility (< 0.05) between different versions of items (invariably associated with low F-ratios) were not considered to offer a reliable basis for interpretation. Similarly, isolated cases involving somewhat larger differences were ignored where these did not achieve statistical significance at the 5% level. However, all effects reaching this level of significance were considered and, where two or more items showed similar trends involving changes of facility of at least 0.05, such trends were examined even when the statistical significance of the individual cases did not reach customarily accepted levels.

A number of trends were identified and seven individual items showed statistically significant effects. In these seven cases, t-tests were used to identify significant differences between the facilities of the standard forms of the items and those of the reduced or augmented versions. Figure 6.3 illustrates the main trends identified and Table 6.13 summarises effects which were statistically significant at the 5% level or better; both have been arranged to show levels of information increasing from left to right.

Table 6.11 Adjusted item facilities from one-way analyses of variance by version, with co-variable MARKRTOT

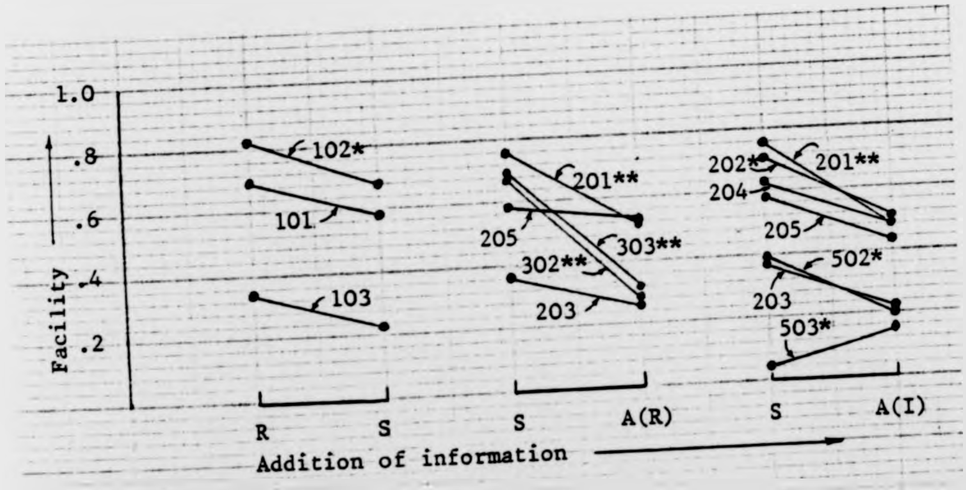
Item number		Information level						
Test	Computer	R	S	A(R)	A(I)	F	p	Notes
1ai	101	.69	.51/.62	.65	–	1.58	.20	R (and to a lesser extent A(R)) tending to raise facility
ii	102*	.82	.60/.76	.73	–	2.68	.05	
b	103	.34	.27/.19	.32	–	1.29	.28	
c	104	.10	.09/.03	.10	–	1.10	.35	No obvious trends
2a	201**	.63	.75	.51	.48	3.66	.01	A(I) (and in items 201, 203 and 205 A(R)) tending to depress facility
b	202*	.71/.73	.71	–	.50	3.19	.03	
c	203	–	.34/.40	.26	.22	2.29	.08	
d	204	–	.65/.61	.60	.48	1.73	.16	
e	205	–	.58/.59	.52	.43	1.28	.28	
3bi	302**	–	.60/.73	.29	.58	8.90	.01	A(R) depressing facility
ii	303**	–	.67/.71	.32	.62	7.73	.01	
ci	304	.66	.63	.68	.71	0.28	.84	Few obvious trends; results of low significance (but A(R) tending to raise facility in items 304 and 405)
cii	305	.65	.69	.65	.70	0.19	.91	
4ci	404	–	.56/.53	.50	.60	0.45	.72	
ii	405	–	.08/.10	.16	.06	1.05	.37	
iii	406	.00/.04	.00	.00	–	1.92	.13	
5b	502	–	.43/.34	.38	.20	2.55	.06	A(I) tending to depress facility
c	503*	.02	.04	.10	.15	3.08	.03	A(I) (and to a lesser extent A(R)) raising facility
6a	601	.43	.32/.40	.46	–	0.97	.41	A(R) tending to raise facility
c	603	.16	.15/.20	.29	–	1.55	.20	
7a	701	–	.36/.42	.33	.33	0.41	.74	No obvious trends; results of low significance
b	702	–	.28/.17	.22	.24	0.72	.54	
c	703	–	.05/.10	.03/.14	–	1.93	.13	

Key:
- R – restricted
- S – standard
- A(R) – augmented (relevant)
- A(I) – augmented (irrelevant)
- F – variance ratio
- p – probability/significance
- * – $p < 0.05$
- ** – $p < 0.01$

Table 6.12 Adjusted item facilities from two-way analyses of variance (item 305 by version and gender; item 502 by version and school) with co-variable MARKRTOT

Item number		Information level				F	p	Notes
Test	Computer	R	S	A(R)	A(I)			
3c(ii)	305	.65	.69	.64	.70	0.25	.86	Non-significant
5b	502*	-	.43/ .34	.38	.20	3.25	.02	A(I) depresses facility

Key: see Table 6.11



Key: see Table 6.11

Figure 6.3 Main trends in the corrected facilities of different versions of items

Table 6.13 Statistically significant differences in corrected facilities of different versions of items

Item no	Versions /	facilities	df	t-statistic
102	R/.82	S/ .68	109	2.12*
201	S/.75	A(R)/.51	113	2.76**
201	S/.75	A(I)/.48	109	3.04**
202	S/.71	A(I)/.50	112	2.43*
302	S/.67	A(R)/.29	109	4.51**
303	S/.69	A(R)/.32	109	4.23**
502	S/.39	A(I)/.20	111	2.42*
503	S/.04	A(I)/.15	109	2.06*

Key: see Table 6.11

Inspection shows at once that, as regards the main trends and for seven out of eight statistically significant effects, a rise in information level was associated with a fall in item facility.[1] This was anticipated where the information added was irrelevant, but was contrary to expectations in the cases in which relevant or even "essential" information was added. It will be convenient to discuss first the cases relating to irrelevant information.

(i) **The influence of augmented irrelevant information.** Six items showed obvious falls in facility associated with the addition of irrelevant information; in three cases this effect was fairly large and was statistically significant at the 5% level or better (Figure 6.3). This was in line with the expectation that such information would tend to have a distracting effect on some students. Table 6.11 also shows that there were eight items for which the addition of irrelevant information made little difference to performance. In all these cases, however, the added data was noticably less substantive than in those

1. A very weak reverse trend, involving a minority of items augmented with relevant information, was also observed and is discussed in sub-section (iii) below.

showing the expected trend. Thus, for example, adding data on the physical state of the reactants in three items, or specifying the nature of a containing vessel in another, did not appear to have a distracting influence. The main trend is exemplified below with reference to particular items. Consideration is also given to one further item (503) which appeared to show a contrary effect at a statistically significant level.

Items 201 and 202 related, in the standard form, to a table showing the number of electrons in the neutral atoms of seven elements, each of which was designated by an arbitary letter of the alphabet. Item 201 asked which two elements belonged to the same group in the periodic table. In the augmented (irrelevant) version, data concerning the mass numbers of the elements had been added to the table and this appeared to distract a significant proportion of the students. In the 30 self-reporting forms examined for this version of the item, five of those who responded to the request to show workings made some reference to mass numbers, and four of these appeared to have selected their answers on this basis. Similar misuse of data relating to mass numbers was noted in students' protocols during the first phase of the study. Although caution is necessary in attempting to draw conclusions from individual results which do not reach significance at the 5% level, it may be noted that similar trends, attributable to the irrelevant mass number data, were observed with items 203, 204 and 205.

Item 202 related to the same data as the previous item and required students to identify the atoms which were isotopes of the same element. This could be done by reference to the numbers of electrons alone but, in the augmented (irrelevant) version, the formula of a compound formed with a familiar element had been provided for each entry in the table.

This significantly depressed the facility of the item. Although it was assumed that this was due to the introduction of irrelevant cuing, no information could be found from the scripts or self-reporting forms either to support or refute this.

The standard version of item 502 simply asked students how they would show that a colourless liquid was water. The addition of an irrelevant "story line" whereby the colourless liquid was obtained after heating urea, made the item significantly more difficult, particularly for students in the lowest achievement tertile. Because of the strong interaction with achievement level, discussion of supplementary evidence relating to these effects from the students' scripts and self-reporting forms will be postponed until Section 6.3.2.

Item 503 showed a contrary effect to that observed in the case of the items discussed above; however the rise in facility involved in this case did not appear to be attributable to the irrelevant material added. In the standard item students were asked to draw the full structural formula of urea, given the contracted form $(NH_2)_2CO$. The information added to the augmented (irrelevant) version was that urea decomposed on heating. It seems impossible to ascribe the raised facility of the item to this and no evidence for any such conclusion could be found in the students' scripts and self-reporting forms. The item was very difficult in all its versions and the elevation of the facility from 0.04 to 0.15 was much smaller than in any of the other significant effects observed. It was significant only because of the very small standard deviation of scores on the item, and this in turn was associated with the omission of the item by many candidates. The effect might therefore be considered as a statistical freak, although an unexpected effect due to context was also considered possible. In the standard version, urea was mentioned for the first time in item 503 itself. However, in the augmented (irrelevant) version, urea was introduced as part of the irrelevant information in the previous item (item 502; testing for water). It might be that students who had realised that the unfamiliarity of urea did not prevent them from attempting the earlier item, might have been less inclined to omit the subsequent part. Some little support for this suggestion was found on examining the students' scripts. While 16 out of 30 students omitted the standard form of the item, only nine out of the 30 omitted the augmented (irrelevant) version. Clearly, whatever the cause, the result was anomalous and it will not be discussed further.

(ii) **The influence of reduced levels of information.** It may be recalled that "reduced" items were produced by the removal of information from "standard" versions. It had been predicted that decreasing the amount of relevant cuing, together with the introduction of some level of ambiguity and the need to make assumptions, would tend to depress item facility. No such effect was observed, most items showing little change of facility as compared to the "standard" versions. In three cases, however, the removal of information was associated with a small rise in facility which reached significance at the 5% level in one case (Figure 6.3). It thus appeared that students were generally insensitive both to the fairly mild levels of ambiguity involved and to the particular cues which had been removed; and indeed that supposedly helpful information in standard items might in fact have a distracting effect for some students and tend to depress the facility of the item. It will be convenient to consider briefly below, first those items which showed a definite, if unexpected, trend before going on to discuss the remainder.

The standard version of items 101 and 102 demanded the name of the anode and cathode products when dilute sulphuric acid was electrolysed between platinum electrodes. Reference to the nature of the electrodes was omitted in the reduced version. This rendered the items noticably easier (at a statistically significant level in the case of item 102). There was no evidence from scripts and self-reporting forms that any student regarded the task as inadequately defined, and clearly the absence of any reference to platinum did not represent the loss of a helpful cue. The most parsimonious interpretation of this finding is that, for the majority of students, reference to platinum electrodes, far from defining an essential part of the chemical situation described in the task, and/or providing an additional cue to the answer, merely represented irrelevant, and therefore distracting, information. Tending to support the latter assertion, it was noted that two students wrote platinum as the answer to the standard version! A similar rather weaker effect, was observed for item 103.

The absence of any trends on the remaining nine reduced items, which were essentially no more and no less difficult than the corresponding standard items, demands comment. In one item (number 304) the removal of an apparently significant cue which was, strictly speaking, redundant, made no difference to performance. Beyond noting that even seemingly relevant cues do not necessarily lead to a rise in facility, little can be said about this case. In four items (numbers 104, 201, 202 and 406) the information removed was a relatively minor constraint included in the standard item to avoid an ambiguity. There was no evidence from the scripts or self-reporting forms that students perceived the ambiguity in the reduced items, nor did it affect their performance. In the remaining four cases (items 305, 503, 601 and 603) the reduced items were both ambiguous and lacking in cues provided in the standard item. There was again no evidence that the ambiguity was perceived, nor that students' performance on these particular items, which covered a range of difficulties, was sensitive to the absence of apparently relevant cues. However, it should also be noted that, in order to avoid the danger of disconcerting and perhaps discouraging students to the detriment of other aspects of the study, the reductions of information in these items tended to be relatively modest. It might therefore be unwise to attempt to generalise from these findings to cases involving more obvious ambiguities.

(iii) **The influence of augmented relevant information.** It had been predicted that increased levels of cuing, in items augmented with relevant information, would lead to a rise in facility. However, although such a trend was apparent in eight out of 22 items, it was only a weak one. Facilities were elevated only modestly and the F-ratios for the items concerned were all low and far from reaching statistically significant levels ($p \geqslant 0.20$). Nine items showed no obvious trends at all and the remaining five a strong reverse trend. Three of the latter items showed effects at statistically significant levels ($p < 0.05$) with large depressions of facility (Figure 6.3). These findings were interpreted in terms of the students' inability to make use of the additional information provided in these cases. Each of these three groups of items are further discussed below with reference to particular examples.

The eight items showing a weak apparent rise in facility associated with the addition of relevant information were numbers 101, 102, 103, 304, 405, 503, 601 and 603.[1] Although the facilities showed only modest

1. Because of the weakness of the effect and to avoid a cluttered diagram, these items were not included in Figure 6.3.

rises in the range 0.05 to 0.10 and were not statistically significant individually, they nevertheless appeared to suggest that some students were able to benefit from certain cues. For example, in items 101 and 102 the addition of the information that the products of the electrolysis of dilute sulphuric acid were gaseous, enabled them to be identified by about 7% more students. Similarly in item 304 the addition of the name of the catalyst apparently enabled about 5% more students to identify the Contact Process. Unfortunately there was little evidence from the students' scripts and self-report forms to show how the additional information was used. Thus in items 601 and 603 students were asked to name the gases formed from given pairs of reactants. Although the facilities of both items rose by about 0.10, there was no written evidence that any student actually used the additional information that one, named reactant was oxidised by the other. In general, although some cases of facilitation were observed, it appeared that only a small minority of students were likely to benefit from the addition of any particular relevant cue.

Support for this interpretation was provided by consideration of the nine items which showed no effects attributable to increased levels of cuing. Inspection of students' scripts and self-reporting forms suggested no clear patterns of behaviour. In some cases students appeared to ignore the additional information. Item 305 asked for the name of the industrial process involving the catalytic reduction of nitrogen to ammonia. Additional information concerning the catalyst appeared to be of little value; students seemed either to know the answer, in which case the added information was redundant, or they did not know it and the added information was of no help. In the 30 scripts and self-reporting forms examined for this item, only two students mentioned the catalyst as providing a cue, conversely one student, having answered the item correctly, expressed doubts since he did not recognise the catalyst and was clearly distracted by the information. However, in some items the information added was certainly utilised even though it failed to have a significant influence on the facility of the item. Item 404 provided an example of this. In the standard version a simple cell was described and students were asked which of the two metals named as electrodes (zinc and copper) would be the positive terminal. The relevant information added in the augmented version, was a listing of eight elements, including zinc and copper, giving their "order in the electrochemical series". While the scripts and self-reporting forms of students who received the standard version made little reference to the electrochemical series (two cases out of 30), the majority of those who received the augmented (relevant) version did so (21 cases out of 30). Of these, however, only four used the extra information effectively, and it may be that these were students who would have used this approach in any case. The majority used the information in a confused manner, almost a third giving an element from one or other extremity of the list as an answer instead of one of the metals actually used in the cell! It seemed clear that, although in this instance many students were attracted to information they were unable to use, they could not have answered the question (except by guessing) in any case.

Five items showed a tendency, very strongly in three cases, for the addition of relevant information to lead to a fall in facility. The standard version of item 201, involving the tabulation of information regarding the number of electrons in certain neutral atoms, has already been described (sub-section (i) above). Students were asked to state which two elements belonged to the same group in the periodic table.

The augmented (relevant) version of the item provided additional information in the form of the formula of a compound formed by each of the elements tabulated. This provided an alternative way of solving the item, or a way of checking the solution obtained by other means. The sample scripts and self-reporting forms examined, suggested that the majority of students who received the standard version solved it by the same S2 (propositional reasoning) strategy employed by the students who participated in the first phase of the study, and that outright guessing was rare. However, a number of those receiving the augmented (relevant) version appeared to have become confused by the data and guessing was more common. Thus, two of those who provided information on their "working" referred rather vaguely to valency and ticked the strategy box representing a guess. Another referred to "combining power" and selected the two elements showing, from their compounds, valencies of one (failing to note that their oxidation numbers were +1 and -1 respectively). Two more mentioned being "confused" and a third referred to "compounds they form" in the part of the self-reporting form asking about difficulties encountered. Similar effects were observed later in the same question with items 203 and 205. Students generally either ignored the additional data or became confused in trying to use it.

Items 302 and 303 can be considered together. The standard versions informed students that two out of four given changes represented oxidation, and asked them to underline the two substances being oxidised. The four changes were:

(i) $Fe^{3+} \longrightarrow Fe^{2+}$ (ii) $H \longrightarrow H^{+}$

(iii) $Cu \longrightarrow Cu^{2+}$ (iv) $Br_2 \longrightarrow 2Br^{-}$

In the augmented (relevant) version, equations were given, thus:

(i) $Fe^{3+} + H \longrightarrow Fe^{2+} + H^{+}$

(ii) $Cu + Br_2 \longrightarrow Cu^{2+} + 2Br^{-}$

The relevant contextual information provided by the equation had been expected to raise the facility of the item in two ways. Firstly by reducing the task from the selection of two amongst four, to the simpler task of selecting one from each of two pairs, and secondly by providing contextual cues. In fact the facility of both items was reduced dramatically from about 0.7 to 0.3, an effect significant at the 1% level. It must therefore be supposed that the additional information represented by the change of format confused rather than helped students. Inspection of the scripts and self-reporting forms provided interesting supplementary evidence. The previous item had asked for a definition of oxidation, and those report forms on which the students' working was shown indicated that definitions in terms of electron exchange, or of oxidation number, were employed in S2 (propositional reasoning) strategies. While examination of 30 scripts relating to the standard version showed no cases of error E-2 (inversion of a proposition during processing), no less than 12 out of 30 scripts relating to the augmented (relevant) version, showed such errors, although in one case the student had subsequently corrected his answer. The E-2 error was tentatively associated, in the first phase of the study, with STM overload and this explanation appears to be confirmed by the present example. Thus while no problems were encountered in applying a simple proposition to changes involving one species only, those students who failed to analyse the more complex situation into its component parts appeared to run a significant risk of being unable to keep sufficient items of information clearly in mind to solve the problem successfully.

6.3.2 **The interaction of information level with student achievement level.** The results of the two-way analyses of variance of item scores (facilities) by information level and performance level (tertile), with MARKRTOT as co-variable, are summarised in Figure 6.4. It had been

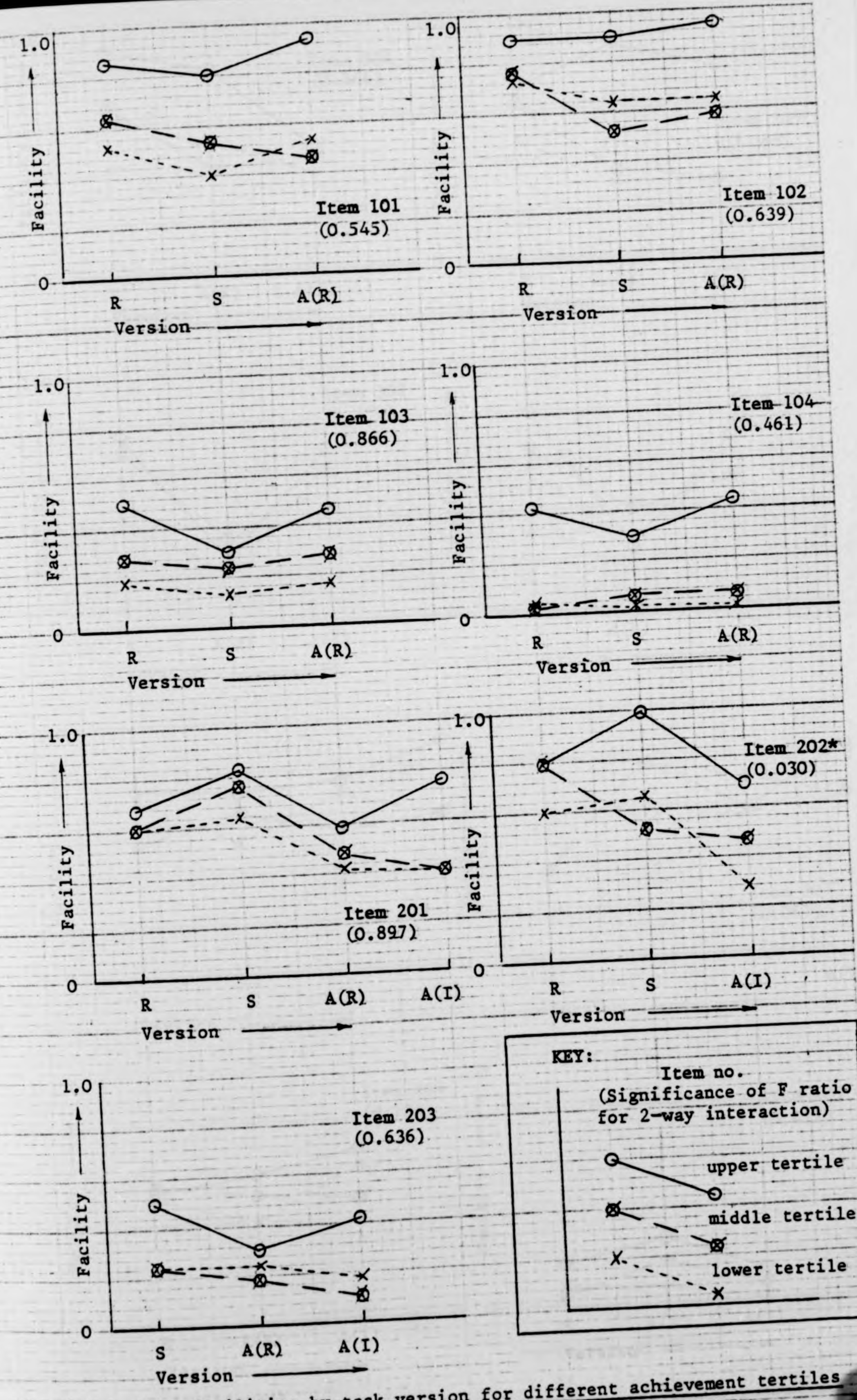


Figure 6.4 Item facilities by task version for different achievement tertiles

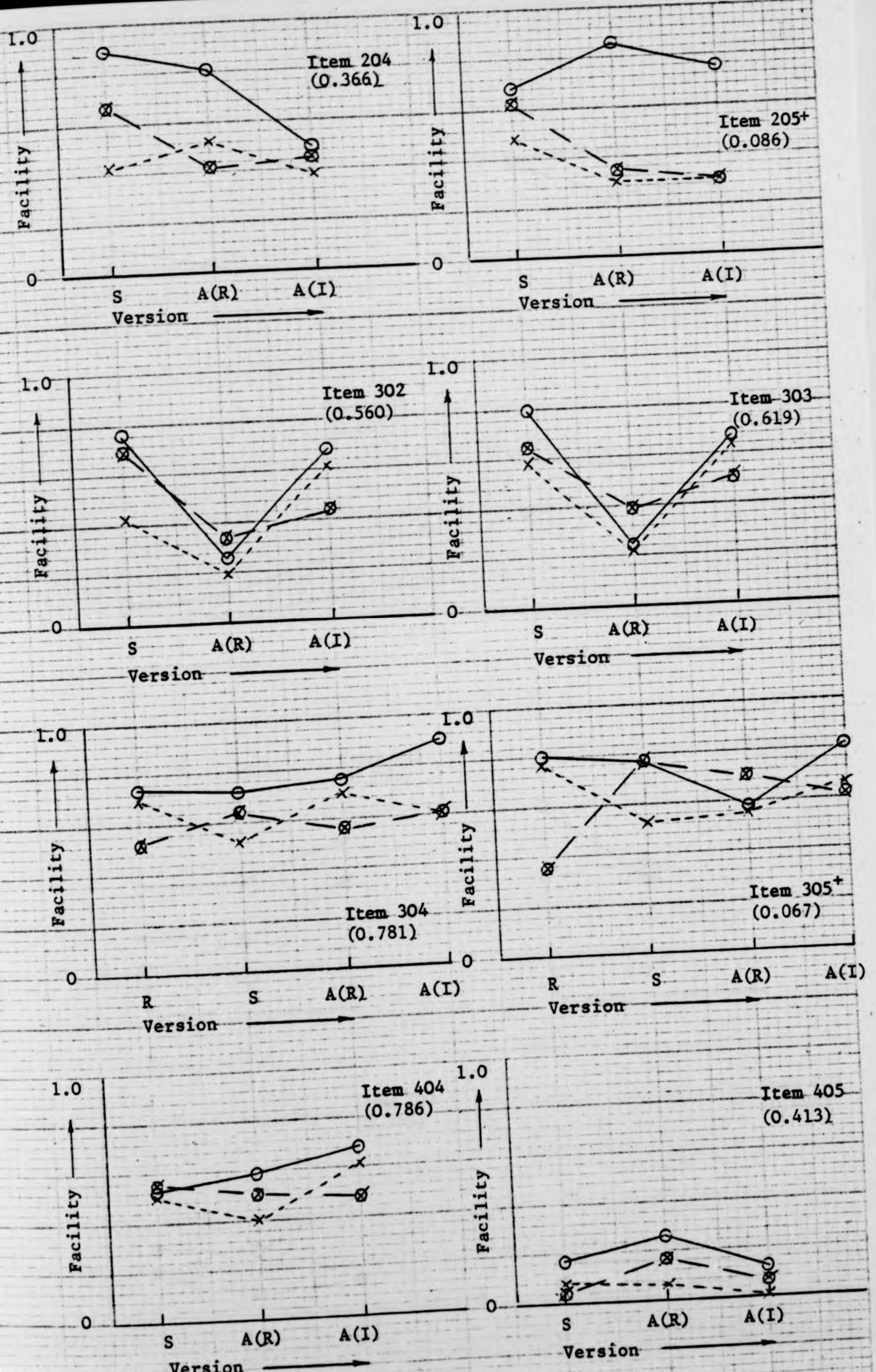


Figure 6.4 (continued)

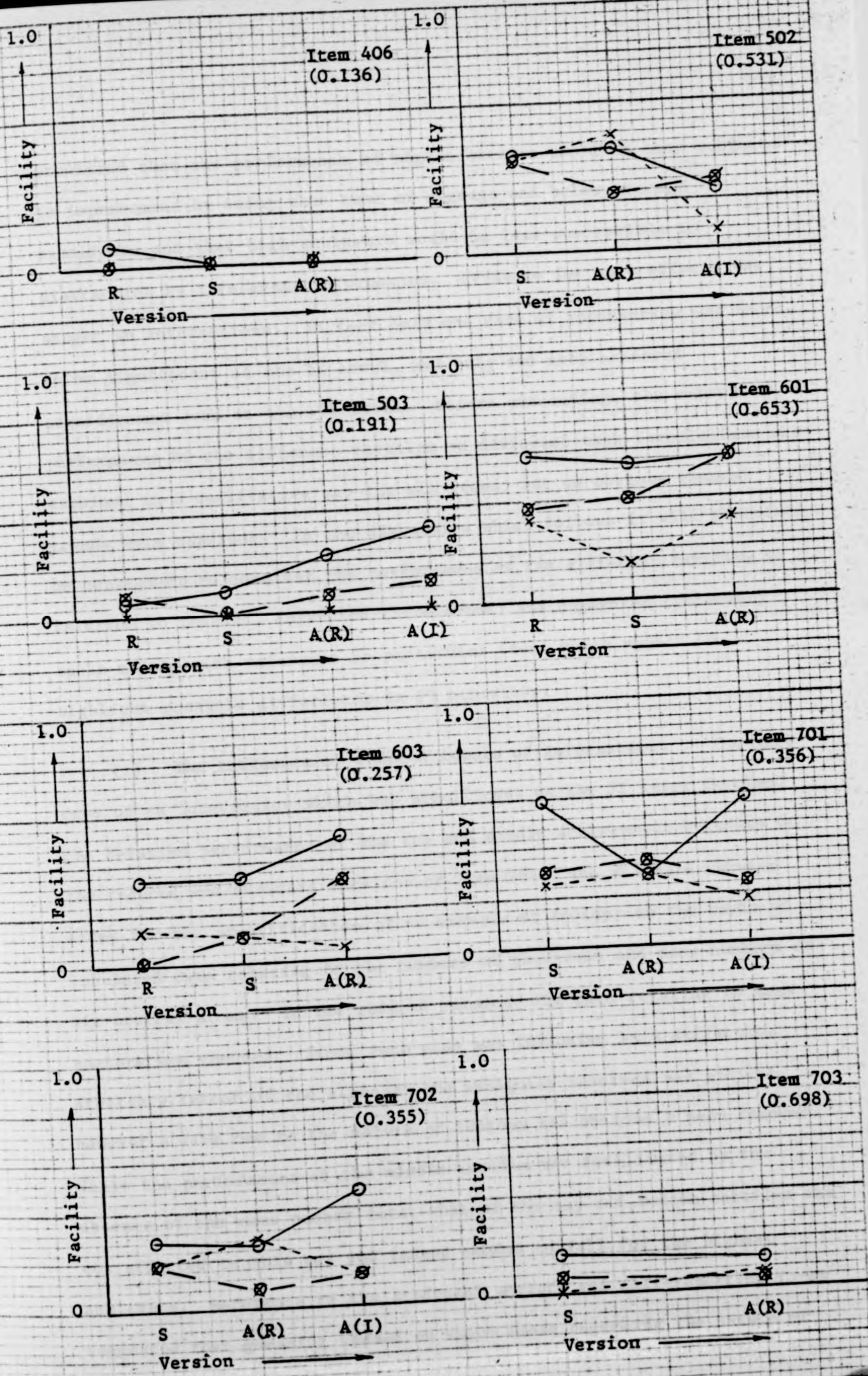


Figure 6.4 (continued)

predicted that the performance of high achievers would be less subject to improvement by cuing than that of average and below average candidates and that high achievers would be less susceptible to distraction by irrelevant information. Evidence for such effects was sought in interactions. In fact only one case of interaction was found to be significant at the 5% level, and only two more achieved significance even at the 10% level. Before proceeding, the relative performance of the different tertiles on different task formulations was examined in a qualitative way for all items, but no obvious general trends were apparent. In the discussion which follows it will therefore be convenient to consider the performance of the different tertiles in relation to reduced, augmented (irrelevant) and augmented (relevant) tasks mainly with reference to particular items which appear to highlight possible differences in PS behaviour.

(i) **The influence of reduced levels of information.** The only case of an interaction which was significant at the 5% level involved the relative performance of the top and middle achievement tertiles on the reduced and standard versions of item 202. The standard version asked for the identification of an example of isotopy on the basis of tabulated data relating to the numbers of electrons and mass numbers of six neutral atoms. In the reduced version, the data relating to mass numbers was removed. Since each atom was indicated by a different, arbitrary letter of the alphabet the ambiguity involved was minimal, however a firm cue to the nature of isotopy had obviously been removed. While the performance of the middle tertile was facilitated by the absence of the mass number data, that of the top and bottom tertiles was not affected in this way and indeed tended to fall for the highest achievers. Reference to the students' scripts and self-reporting forms suggested that guessing (based on vague ideas regarding the nature of isotopy and attempts to seek regularities in the data given) was relatively frequent for both bottom and middle tertiles, irrespective of version. Obviously this approach was more likely to lead to the correct answer in the reduced version. Indeed, in the standard version several students, having failed to analyse the nature of the task in terms of the information provided, subtracted numbers of electrons from mass numbers in a pointless attempt to identify atoms with different numbers of neutrons. Students who appeared to know the definition (including all those in the top group) however, tended to behave differently according to their achievement tertile. Out of the 10 students from the bottom tertile (reduced version) whose scripts and self-reporting forms were examined, four stated that they were unable to relate their definition to the data provided and omitted the item. This would obviously tend to counteract the facilitation associated with those who were prepared to guess. None of the sample of middle tertile scripts and self-reporting forms examined made any reference to the limited data provided in the reduced version, however three students representing the top tertile stated that the information was inadequate, although only one omitted the item (after crossing out the correct answer). To summarise, three effects appeared to be operating:

(a) guessing and solutions based on general intuitions regarding the nature of isotopy were favoured by the version offering least data;

(b) the top tertile was the most likely to perceive the ambiguity in the reduced version and their performance may have been slightly depressed by this;

(c) the bottom tertile were least able to relate their definitions to the given data.

It should also be mentioned that there was no evidence from the scripts and self-reporting forms, or from the performance data, that reference to mass numbers in the standard version either did or did not help to cue the correct definition of isotopy.

Item 305 showed an interaction, significant at the 10% level, involving the performance of the different tertiles on the standard and reduced versions of the task. The standard form of the item asked students to name the industrial process involving the catalytic reduction of nitrogen to ammonia. In the reduced version the words "to ammonia" had been omitted. The loss of information did not influence the performance of the top tertile, but depressed that of the middle tertile while raising the performance of the bottom group. No students appeared to perceive any ambiguity in the item and inspection of the scripts and self-reporting forms provided limited help. The cue "ammonia ——→ Haber process" appeared to be strong for all groups on the standard version, however while its absence was critical to many in the middle tertile, this was not the case for the high achievers. It might be presumed that, for the latter group, cognitive structures relating to the Haber process were sufficiently well developed for "the catalytic reduction of nitrogen" to be an adequate cue. The enhanced performance of the lowest tertile on the reduced version could not be explained. However, the interactions in this item were of only limited statistical significance and thus all interpretations should be treated with caution.

Interesting, though not statistically significant, effects were also noted in relation to item 102. The enhanced facility of the reduced verion of this item was discussed in the previous section. Inspection of Figure 6.4 suggested that this facilitation was due entirely to a rise in the performance of the middle and lower tertiles. It will be recalled that item 102 related to the electrolysis of dilute sulphuric acid and that reference to the nature of the electrodes (platinum) had been removed in the reduced version. It appeared that, as predicted, high achievers were not distracted by the reference to platinum in the standard version, while average and below average candidates were.

(ii) **The influence of augmented irrelevant information.** Figure 6.4 illustrates interesting, though not statistically significant, trends on two items. It will be recalled that the standard version of item 201 asked students to identify two elements in the same group of the periodic table on the basis of data relating to numbers of electrons in neutral atoms. Data relating to mass number was added in the augmented (irrelevant) version. A significant fall in facility associated with this change was ascribed to the distracting effect of the additional data. Inspection of Figure 6.4 shows that this distraction was most evident for the two bottom tertiles and had little effect on the higher achievers. All the five cases reported earlier, in which evidence of distraction by the mass number data was found in students' scripts and self-reporting forms, occurred in the middle and bottom tertiles.

Findings were similar in relation to item 502, the standard version of which asked students how they would show that a colourless liquid was water. The addition of an irrelevant "storyline" in which the colourless liquid was obtained on heating urea, caused a depression in the facility of the item. Reference to Figure 6.4 shows that, in this case, the distracting effect of the irrelevant data was largely confined to students in the lowest tertile. Four students, out of the ten in this tertile whose scripts and self-reporting forms were examined, quoted unfamiliarity with urea as causing difficulty and two of them omitted the item. No similar comments were made by students in the other tertiles.

(iii) **The influence of augmented relevant information.** Four items will be discussed in relation to the influence of augmented levels of relevant information on the success of the different achievement tertiles.

Three of the items considered showed no evidence of significant interaction. It will be recalled that the standard version of items 302 and 303 asked students to identify, in four symbolically represented changes involving ions, two substances which were being oxidised. The facility was dramatically depressed when pairs of changes were combined as equations, the depression being attributed to STM overload rather than to changes in the information level as such. Reference to Figure 6.4 shows that this effect applied almost equally to students of all achievement levels. Item 201, relating to the identification of two elements as belonging to the same group of the periodic table, showed a similar pattern of behaviour. The addition of examples of formulae to the standard data on numbers of electrons, depressed the facility of all tertiles equally. Evidence, from the students' scripts and self-reporting forms, of some confusion concerning the use of information from the formulae has already been reported. Similar confusion was often associated, in students' protocols obtained during the first phase of the study, with STM overload. Although direct evidence was lacking, it seemed plausible that students seeking to generate and retain valencies from parallel, but quite different, sets of data, might well encounter such a problem. If this is assumed, then a consistent pattern emerges. Thus significant depressions in facility which are associated with distracting information, tend mostly to affect students in the middle and lower achievement ranges, while significant depressions associated with STM overload resulting from an excess of more relevant information, tend to affect students of all achievement ranges equally.

Finally, reference must be made to item 205 which showed an interaction, significant at the 10% level, involving the facilities of the standard and augmented (relevant) versions. In the standard version, students were asked to derive the formula of a simple, binary compound on the basis of the numbers of electrons in each of the two neutral atoms concerned. In the augmented (relevant) version, the formulae of their compounds with sodium and with chlorine were provided as additional data. Figure 6.4 suggests that the performance of the top achievement tertile was enhanced while that of the middle and lower tertiles was depressed. Reference to the scripts and self-reporting forms suggests that most students in the top tertile ignored the added data (although in one out of the ten cases examined, the student concerned ignored the electron data and worked only from the formulae). Conversely, many of those in the lower achievement groups attempted to use information from both sources and became confused. Since the majority seemed unable to make effective use of any of the information, this should probably be regarded as a case of distracting information rather than of STM overload, although the distinction may be somewhat blurred here.

6.3.3 Other significant interactions involving information level.

For the sake of completeness only, two cases of significant interactions involving information level with, respectively, gender and school, are illustrated in Figure 6.5 and are referred to very briefly below.

Item 305 was the only item to show a significant interaction of information level with gender. The interaction was associated with the augmented (relevant) version and showed that naming the catalyst as "finely divided iron and promoters" made it more difficult for boys but not for girls, to recognise the Haber process. The students' scripts and self-reporting forms provided no help in interpreting this finding which seems to imply that girls tended to be familiar with the nature of the catalyst, while boys tended not to be and were distracted by it.

Item 502 was the only item to show a significant interaction of information level with school. Reference to Figure 6.5 shows that students at school C found the augmented (relevant) version of the item more difficult than the standard version, while those in the other three groups found it easier. In the absence of detailed information regarding the teaching programmes at each school, no attempt was made to interpret this finding

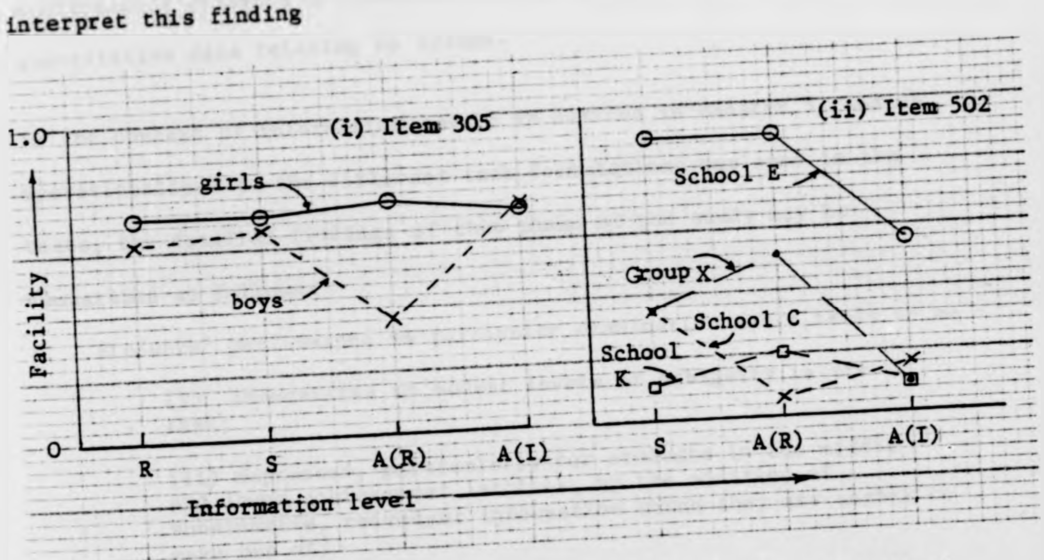


Figure 6.5 Significant interactions involving information level and (i) gender, (ii) school

6.4 A discussion of the findings

An obvious feature of the analyses reported above was the relative scarcity of statistically significant effects. Out of 49 possible main effects due to information level, only eight were "significant" at the 5% level. A further seven effects reflected the same trends rather less strongly and eight more showed an additional weak trend which was not demonstrated by any statistically significant cases. As regards the interaction between information and achievement levels, the proportion of statistically significant effects was still lower and general trends were hard to find. However, the relative scarcity of effects was itself important, showing PS performance in the context of examination tasks to be fairly resistant to modest changes in the level of information provided across all achievement levels. Against this background, a number of conclusions are elaborated and discussed below, the limited data available from students' scripts and self-reporting forms providing confirmatory evidence for certain interpretations of the more quantitative data relating to trends.

In the context of information levels as defined in Chapter 5, and as operationalised in the different task formulations employed in the tests, the detailed findings of this phase of the study may be summarised as follows:

Students' performance on particular examination tasks tends to be -

(i) insensitive to modest levels of ambiguity in the task;

(ii) depressed, particularly for students in the middle and lower achievement tertiles, by the addition of substantive, redundant information which they are unable to make use of;

(iii) only weakly enhanced by increased levels of relevant information;

(iv) depressed, in tasks demanding propositional reasoning, by the addition of redundant information even when they are able to make use of it.

These findings are elaborated and discussed below and some of their implications are considered.

6.4.1 **Ambiguity in tasks.** There was no evidence that the performance of students at any achievement level was significantly influenced by the ambiguities introduced in 11 out of the 12 reduced versions of items. In only one case was there firm evidence that any ambiguity was perceived. The students obviously would not have anticipated ambiguity in examination items, and their perceptions may have been influenced by such expectations. However, this finding is consistent with that of

Ward (1982), relating to students' interpretations of observational data in the context of A-level inorganic analysis, when he reported that many seemed to be unaware of the limited nature of the conclusions which could be drawn from a particular observation. The common feature of these findings is reflected in the apparent tendency of students to accept certain "obvious" or "expected" conclusions uncritically, without proper reference to the adequacy of the evidence. While some caution must be exercised in attempting to generalise from PS in examination situations to PS in a wider context (particularly in the present case where only modest levels of ambiguity were involved) failure to perceive or to take into account the limitations of available data represents a serious constraint on effective everyday PS in chemistry, whether at the laboratory bench or elsewhere.

6.4.2 **The distracting effect of non-usable information.** The addition of substantive information to an item tended to depress performance if the information was unfamiliar to the students or if, owing to inadequacies in their own knowledge or skills, they were unable to utilise it. Although some of this information was considered "relevant", or even essential to the task, by the author, it was often not useful, and thus "irrelevant", to the students who rarely made judgements regarding the information necessary to solve the item and tended to treat all data as being of equal importance. Thus unfamiliar information was perceived as discouraging and redundant information of a more usable kind was distracting or confusing. Although it is possible that this may have reflected, to some extent, students' expectations regarding the information provided in examination items, it was notable that the tendency towards what was identified as "serial processing" in the first phase of the study and as "simultaneous scanning" by Ward, was apparently less prevalent for students in the top achievement tertile. In some cases this might be attributed to a greater ability to discern the demands of the task and thus to perceive the redundancy or irrelevance of some of the data; in others, possibly to a greater capacity to utilise redundant information of a more relevant kind. The former case clearly echoes Krutetskii's (1976) findings regarding similar differences between mathematically gifted and average children in analysing the structure of a problem and thus in discriminating between relevant and irrelevant information. In real-world PS situations, where only a proportion of the information available is likely to be relevant to the particular task in hand, the ability to discriminate what is needed from what is not, is likely to be a prerequisite for adequate performance.

6.4.3 **The effects of additional cuing.** It had been predicted that the facility of an item would be raised by the cuing provided by additional, relevant information. Although a number of items showed a weak trend in this direction, in no case did this approach statistically significant levels. In a number of items the information proved to be "irrelevant" to many students and resulted instead in a lowering of facility. It was concluded that, if the necessary information structures in LTM were not cued by the "standard" (or in most cases a "reduced") form of the task, then further cuing was likely to be of only limited effectiveness, and might even be distracting.

It may be instructive to consider further the apparent interaction ($0.05 < p < 0.10$) involving the middle achievement tertile[1] on the simple recall item 305 (Section 6.3.2, (i)). For the other tertiles,

---

1. The "marker" items discriminated well at high achievement levels but only moderately so for low achievers. The "middle" tertile proved to be much the weakest group on this particular item.

"industrial process", "nitrogen" and "catalytic reduction" were sufficient to cue "Haber" for nearly 80% of students, and the addition of "ammonia" failed to activate any greater proportion. However, for the "middle" tertile the limited information version cued the correct response for only about 30% of the students, while the addition of "ammonia" raised this to 80% giving a strong impression of a threshold-like effect. This might imply that the likelihood of activating a particular item of information in LTM was related simply to the number of (relevant) cues provided. However, it seems more helpful to interpret the activation of appropriate information in terms of the extent to which the particular concepts involved in solving a problem are interrelated within an individual's semantic network. Thus, while students with well integrated structures might be able to correctly identify the Haber Process from several different starting points, activation of the concept for many in the "middle" tertile appeared unlikely in the absence of the salient cue "ammonia".

Although the item quoted represents a contrary example, in general, when a particular task failed to activate the necessary areas in LTM, supplementary cuing from additional information also failed to do so. Reference to the widespread incidence of this phenomenon, as well as to particular cases, suggested that the absence of "salient cues" was very rarely a tenable explanation and it appears that many students tend to be inflexible in their perceptions of a task and seem unable to approach it from more than one direction. To the extent that this is the case, it has serious implications for PS outside the examination hall. It calls to mind the notions of "set" and "functional fixedness" associated with much of the classical research literature of the Gestalt school, as well as some of the subjects in Ward's much more recent research who clung to early, erroneous, conclusions despite subsequent conflicting evidence.

6.4.4 Short-term memory problems. While in the great majority of cases the addition to an item of redundant information of a readily usable kind produced at best a small improvement in PS performance, in three items solved by propositional reasoning a sharp fall in facility was observed and appeared to be associated with STM overload. In these cases it appeared that many students had failed to analyse the demands of the task sufficiently clearly to identify the information which they actually needed. Thus attempts to process and combine all the available information indiscriminately, led to typical STM overload errors of the kind classified as E2 and E-2 in the first phase of the study. It was notable that students of all achievement tertiles were equally vulnerable. This finding reinforced earlier conclusions that the failure of students to identify clearly the demands of the task must place serious limitations on their performance, particularly in situations where the information level is less tightly controlled than is usually the case in examinations.

6.4.5 A summary of conclusions. The immediate objective of this part of the study was to investigate the influence, on the students' PS performance, of the level of information provided in the task. Findings relating to this have been reported above. Underlying this objective was the hope of obtaining further insights into the nature of students' PS processes, with particular reference to bridging the gap between PS behaviour in examinations and in more realistic settings where information levels are less controlled.

The students whose PS behaviour was investigated were all being entered by their schools for the O-level chemistry examinations within a few weeks of participating in the study. While a small proportion were considered borderline cases and were also entered for the CSE examination, the majority were confidently expected by their teachers to pass. Although all the findings, except that concerning vulnerability to STM overload, may apply less strongly to the highest achievement tertile, there appears to be some cause for concern that successful examination candidates may be ill-equipped to tackle realistic chemical problems. Outside examinations it is rare for the level of information available to be properly matched to the task. The case of elementary inorganic analysis, which was selected by Ward for his study of students' interpretations of observational data, is a good example. Some of the information obtained during such an analysis may be ambiguous, irrelevant or redundant. The findings reported above suggest that students often fail to recognise ambiguity and thus the limited nature of the conclusions possible in such circumstances. They also tend to be undiscriminating in handling data which goes beyond that actually demanded by the task. Where they are unable to utilise such data this tends to lead to discouragement or confusion and a lowering of performance, while excessive data of a more familiar kind may lead to processing errors associated with STM overload, or at best to nugatory effort and lowered efficiency. Finally, it is suggested that students may tend to be inflexible in their perceptions of the nature of a task and that this may render them insensitive to cuing from relevant data. Insofar as their initial perceptions may be inadequate, or even incorrect, this represents a barrier to successful performance in whatever context.

In summary, it must be concluded that most O-level candidates fail to analyse the demands of chemistry tasks and tend to process available information indiscriminately. While it is obviously necessary to exercise caution in generalising results from a paper-and-pencil test, it also seems reasonable to conclude that, in realistic situations, students may not be able to equal their examination performance owing to the imperfect match between the level of available information and the demands of the task.

CHAPTER 7

THE RELATIONSHIP BETWEEN PROBLEM-SOLVING PERFORMANCE AND COGNITIVE STRUCTURE

The analyses described in this chapter were carried out to investigate possible associations between students' PS performance and their cognitive structures. Following an account of the processing of the data from the word-association tests, an exploratory study of results relating to the test population as a whole (and including an empirical evaluation of certain methodological procedures involved in cognitive mapping) is reported. The account continues with comparisons of the mean maps of groups of students of high and low achievement on relevant sets of items. A number of short case studies are then presented in which particular PS episodes are related to the cognitive maps of individuals. The findings are discussed in the final section of the chapter.

7.1 The processing of data from the word association tests

In order to assess the demands of any analysis, the characteristics of the data available, and of the information required must be determined. The raw data obtained from each of the participating students, consisted of lists of responses to each of 27 stimulus words. The principal requirement was to present this information in the form of the cognitive maps of students, and groups of students, reflecting the semantic organisation in LTM of the concepts represented by the stimulus words. The principles involved in converting lists of associative responses into half-matrices of relatedness coefficients, as a preliminary to generating cognitive maps, were discussed in Chapter 5 and the practical details are reported below. Consideration was also given to the desirability of obtaining supplementary information of a non-structural kind from the data. The limitations of analyses based on the classification of responses were discussed earlier. In particular, the a priori attachment of special significance to "stimulus-list responses" (Johnson, 1967) was considered inappropriate in the absence of any unique structure linking the concepts involved in this study. It was, however, decided that the interpretation of differences between the cognitive maps of selected groups of students could be illuminated by comparisons of the "associative meanings" characteristic of each group. This involved being able to generate, for each group, cumulative frequencies of all responses to each stimulus word. The encoding of the raw data and its processing to generate relatedness matrices and cumulative response lists for any group of students is described below.

7.1.1 **The encoding of data from the word association tests.**

Students' responses to the WAT were encoded for computer analysis using a three digit code which was built up as the encoding proceeded. The final lexicon contained 954 different words and a total of 35,371 responses was encoded. A small proportion of responses demanded special decisions, namely those representing inappropriate responses and those relating to questions of interpretation. Each is considered separately below.

(i) **Inappropriate responses.** These accounted, in total, for less than ½% of all responses and corresponded to problems identified by Noble (1952). Following his precedent, which has since become normal practice, responses were ignored (not encoded) if they were illegible, repetitions of the stimulus word, repetitions of a previous response to the same word, or indicative of a failure of set (that is to say, of a student having moved outside the set of ideas under consideration). The latter will be discussed further under three headings: Failure of contextual set, False strategies, and Alliterative and facetious responses.

(a) **Failure of contextual set.** No formal constraint had been placed on students' responses and no prior judgement had been made regarding the treatment of associations to a word in a non-chemical context. In practice, such associations were found to be extremely (and unexpectedly) rare. This probably reflected the strong constraints implied by the test situation and the highly selective nature of the test population. It was soon apparent that such responses, which were generally idiosyncratic, would be of little statistical or interpretive significance and, in particular, would contribute nothing to the associative overlap between concepts. Where such responses occurred at the end of a list of associates, they were therefore simply ignored. However, where such a response was followed by one or more responses to the stimulus word in its chemical context, an arbitrary code (999) was used so as to preserve the rank order of the subsequent responses. Where, very rarely, a particular student employed a mid-list context-failure response more than once, new codes (998, 997, etc) were employed to avoid the computation of overlaps between response hierarchies where these did not in fact exist. The few cases where there was reasonable doubt as to whether or not a word was being responded to in a chemical or a non-chemical context were always treated as being the former and were encoded normally.

(b) **False strategies.** A few cases of "chaining" were observed in which students, instead of continuing to produce associates to the stimulus word, produced associates to their own most recent response. The identification of such cases obviously involved some degree of subjective interpretation. This was also the case where students employed a conscious response strategy instead of instead of responding spontaneously. For example, after responding normally to ELECTRODE a number of times, one student listed a series of metals, any of which might have been employed as an electrode but which, apart from the first two, were most unlikely to have been encountered in that role. In doubtful cases, responses were always assumed to be normal and were encoded accordingly. However, where chaining or other failures of set were clearly involved, the third and subsequent items in such a sub-list were ignored (the first member being regarded as a normal associate and the second, as this invariably seemed quite possible as a normal associate, being given the "benefit of the doubt").

(c) **Alliterative and facetious responses.** Noble recommends that alliterative and facetious responses should be ignored. For example, a response such as "urination" to the stimulus word OXIDATION might be classified as both and seems most unlikely to relate to any semantic features of the student's concept of oxidation. Responses which were clearly of either sort were very rare and were ignored, although doubtful cases were encoded normally.

(ii) **Questions of interpretation.** Decisions that had to be made to resolve difficulties of interpretation involved less than 5% of responses and will be considered under the headings Ambiguity, Synonymy and Multiple-word Responses. In discussing the first two of these categories, reference will be made to Chomsky's linguistic notions of surface-structure and deep-structure (see for example Greene, 1971).

(a) **Ambiguity.** This occurs when a single verbal response (surface-structure) maps on to more than one concept (deep-structure). In practice, ambiguous responses were rare and easily dealt with. Generally the context provided by the test was sufficient to allow the unambiguous interpretation of potentially problematic cases. Thus there was little likelihood of interpreting the response "arms", to the stimulus word VALENCY, as referring to weaponry. In rare cases, where there was true ambiguity, as in the response "plate" to the stimulus ELECTROLYSIS (did the student have in mind a plate of metal as an electrode, or the verb to (electro)plate?) the simple solution adopted was to eschew interpretation and to assign a single code. This could only influence the pattern of associative overlap if the same word was used, with different connotations, by the same student in responding to a different stimulus word and it is very doubtful if this ever occurred.

(b) **Synonymy.** This occurs when a single idea or concept (in deep-structure) maps on to more than one verbal response (in surface structure). Cases represented by merely symbolic or grammatical variation were ignored. Thus "hydrogen", "H_2" and "H" were encoded as identical responses, as were "atom", "atoms", and "atomic" (providing the latter was not part of a multiple-word response). In conformity with this decision, occasional responses such as "atomic" to ATOM and "+" to POSITIVE were ignored. Somewhat more difficult to deal with was the occasional use of entirely different words for the same concept. All technical terms were encoded differently when different words were used on the grounds that any synonyms might not necessarily be apparent to the students. However, in a few obvious cases involving non-technical language, a single code was used to cover more than one word. Thus "loss", "lose", "give away" and "donate" were encountered as responses to OXIDATION. Generally speaking, any particular student used one version of these terms consistently in all his responses, in which case the coding system adopted was irrelevant insofar as the calculation of interconcept proximities was concerned. However, it seemed probable that a small minority who were less consistent were nevertheless referring to the same concept when associating "gaining" with OXIDATION and "accepting" with REDUCTION. The examples quoted were much the most numerous, but a few other cases called for similar treatment and were reflected as synonyms in the lexicon.

(c) **Multiple-word responses.** The instructions given to students allowed multiple-word responses where these represented a single idea, and examples such as "sulphuric acid" and "Avogadro's number" were given. However, they were also asked to avoid explanatory phrases. In practice, most multiple-word responses presented no difficulty in that they represented a clearly distinctive concept as in the above examples. Thus the response "energy level" (to the stimulus word ELECTRON) was regarded as conceptually distinct from "energy" (as a response to ATOM) and "level" (as a response to ELECTROLYTE) so that these three responses would make no contribution to the calculation of the semantic proximity of the stimulus words concerned. In some cases, however, the situation was less clear-cut. For example in "beta-ray" and "beta-particle" (separately encountered in response to ELECTRON) the elements making up these compound words retain more of their original character; thus the first could be regarded as overlapping semantically with "ray" (in response to CATHODE) and the second with "particle" (in response to ION). Such decisions are essentially arbitrary but in the present study were made on the basis of the author's judgement of the probable distinctiveness of the multiple-word concept, from the separate concepts represented by its parts, in the eyes of an O-level student. Where it was judged that the separate elements largely retained their original character in the compound word, separate codes corresponding to each element were employed. Where, on the other hand, the compound was judged to take on a more independent identity, a single code was preferred. It should be noted that the proportion of responses for which such arbitary decisions had to be made was very small.

A similar approach dealt successfully with the few explanatory phrases which appeared despite instructions to the contrary. Such phrases were broken down into successive elements which were coded separately, words expressing relationships between concepts being ignored as they seldom otherwise appeared. Thus, in response to NEGATIVE such occasional phrases as "opposite of positive" and "not positive" were encoded identically with the more usual response "positive".

Following encoding, the data were punched onto cards for computer processing as described in the next section and in Appendix C.

7.1.2 **The calculation of relatedness matrices.** A measure of the semantic proximity of a pair of stimulus words for a given student can be derived in the form of a "relatedness coefficient", from the overlap of his lists of associative responses to the two words. The calculation of such coefficients, as described by Garskoff and Houston (1963), was discussed in some detail in Section 5.2.6. Using this method, half-matrices of relatedness coefficients were calculated for each student using a short computer program which was written for this purpose.[1]

1. A number of programs which calculate relatedness matrices have been published (Mead, 1972; Shavelson and Lee, 1975; Noval, 1976), however these were not immediately available. It was therefore more convenient to develop an original program which could be tailored to the particular demands of the present study. Annotated copies of program listings will be found in Appendix C.

Subsidiary programs enabled the matrix of any particular student, or the mean matrix of any selected group of students (calculated by averaging the entries in corresponding cells in individual matrices), to be obtained as a printout or as an input to the CLUSTAN 1B suite of programs (Wishart, 1979) for cluster analysis procedures. The programs also calculated mean lengths of response lists and mean relatedness coefficients for each student or for any group of students, together with a frequency count of all responses to each stimulus word for any group of students.

7.2 A preliminary analysis of the mean data for the test population including a comparison of cluster analysis methods

Before proceeding to comparisons between students of high and low performance, a preliminary analysis was carried out using data relating to the test population as a whole. As well as providing an overall picture against which particular groups and individuals might be compared, this provided an opportunity for evaluating alternative methods of cluster analysis. In the sections below, a brief account of the associative meanings of the stimulus words for the test population is followed by a comparative discussion of the population maps derived using selected hierarchical and non-hierarchical clustering methods.

Two preliminary decisions must first be reported. The first concerns the value assigned to Garskoff and Houston's parameter "p" when calculating the relatedness half-matrices (see Section 5.2.6). The importance, in studies involving the mapping of cognitive structures, of taking account of the primacy of a student's associative responses has generally been reflected by using a value of one for "p" and this value was therefore adopted for the present study. Although a case can be made for using a higher value on the basis of Garskoff and Houston's criteria relating to "steep" response hierarchies and low, rather than high, creativity, the rationale for such a decision was not considered persuasive. Little work has been reported using values other than one and a preliminary comparison of hierarchical maps derived from population mean matrices calculated using p = 1 and p = 2, respectively, showed only minor differences and certainly no clear advantage for the latter. The second decision to be reported was also foreshadowed in Chapter 5 (Section 5.2.7). In the absence of any a priori expectation of a dimensional solution, factor analysis and multi-dimensional scaling interpretations of the data were not sought.

7.2.1 **The associative meaning of the stimulus words.** The associative meaning of any stimulus word is defined by the responses which it elicits in a WAT (Deese, 1962). Table 7.1 lists the most frequent responses to each of the 27 stimulus words employed in the study, including all those used by at least 10% of the students. In compiling the table no account was taken of the primacy of responses. The words were arranged in order of their "meaningfulness"; that is, in the order of the mean number of associative responses elicited by each (Noble, 1952). In the context of the mapping study these data were of general interest only; however a brief comment will not be out of place.

The "meaningfulness" of the different stimulus words showed a relatively wide range, ELECTRIC CURRENT at one extreme attracting almost twice the mean number of responses as ELECTROCHEMICAL SERIES at the other. Although any number of particular observations might be made (Why, for example, was CATHODE more meaningful than ANODE?) no significant generalisations were apparent.

Table 7.1 The most frequent associative responses to each of the 27 stimulus words employed in the study

Stimulus word	Mean no of responses	Most frequent responses (% of respondents)
ELECTRIC CURRENT	6.54	electron (47), electrolysis (38), amp, amphere (32), charge (31), volts, voltage (25), flow (23), negative (22), positive (20), battery (18), wire (13), ion (13), electrode (13), anode (12), electrolyte (12), coulomb (11), ammeter (10), cathode (10)
CATHODE	5.93	anode (87), electrolysis (71), , negative (69), electrode (41), positive (25), electricity (21), electron (19), copper (18), electrolyte (18), ion (17), cation (14), hydrogen (13), solution (10)
ELECTRODE	5.62	cathode (71), anode (66), , electrolysis (58), negative (31), positive (31), electricity (30), copper (24), platinum (23), electrolyte (21), carbon (21), graphite (18), electron (14), ion (14), current (11)
ELECTRON	5.60	negative (62), proton (60), neutron (51), atom (39), shell/orbit (22), ion (21), electrolysis (19), electricity (17), charge (16), nucleus (11), positive (11)
PERIODIC TABLE	5.60	group (69), element (54), period (36), halogen (27), atomic number (18), gas (18)*, metal (17), symbol (14), number (13), valency (13), electron (12), alkali metal (11), atomic mass (10) *[inert (6) + noble (6) + rare (4) gas = (14)]
ELECTROLYSIS	5.56	cathode (62), anode (58), , electricity (41), ion (36), electrode (35), electrolyte (25), copper sulphate (16), solution (14), water (14), current (14), positive (12), copper (12), negative (11)
ATOM	5.47	proton (49), electron (48), neutron (46), molecule (35), element (34), nucleus (28), ion (24), tiny, small, minute (23), atomic mass (20), (atomic) structure (18), atomic number (14), orbit (12), bomb (11), particle (10)

Table 7.1 (continued)

Stimulus word	Mean no of responses	Most frequent responses (% of respondents)
ION	5.38	positive (64), negative (63), electron (40), atom (28), eletrolysis (27), charge (25), bond/-ing (16) [ionic- (bond) 12)], metal (11), gain etc (10), anion (10), cation (10), cathode (10)
CHARGE	5.28	positive (83), negative (82), electricity (38), ion (33), electron (32), electrolysis (19), current (16), proton (14), volts/-age (11)
PROTON	5.27	neutron (88), electron (73), positive (60), nucleus (51), atom (50), , atomic number (20), atomic mass (18), charge (16)
ANODE	5.17	cathode (86), electrolysis (77), positive (68), electrode (25), ion (21), negative (19), anion (15), electricity (15), copper (14), oxygen (12), electron (12), carbon (11)
GROUP	4.70	periodic table (68), halogen (40), period (18), element (25), metal (20), number (15), gas (15)*, valency (13), one/1 (11), similar (10) [inert (5) + noble (6) + rare (3) gas = (14)]
POSITIVE	4.62	negative (87), anode (46), proton (37), ion (36), charge (24), cathode (16), electron (13), electrolysis (13), metal (11), hydrogen (11)
ELECTROLYTE	4.60	electrolysis (61), solution (30), electricity (30), cathode (24), anode (22), liquid (21), ion (17), conduction etc (17), copper sulphate (17), electrode (14), electron (13), current (12), molten (12)
OXIDATION	4.54	oxygen (79), reduction (73), electron (44), loss (35), gain (31), redox (12), reaction (10)
VALENCY	4.40	number (37), electron (36), bond/-ing (22), periodic table (21), element (19), group (16), one/1 (12), formula (11), ion (11), two/2 (10), equation (10), combining-power (10)
ELECTRONIC CONFIGURATION	4.22	electron (60), orbit etc (39), atom (19), energy-level (19), proton (17), ion (17), number (14), bond/-ing (12), neutron (12), structure (11), K/L/M etc (10)

Table 7.1 (continued)

Stimulus word	Mean no of responses	Most frequent responses (% of respondents)
ELEMENT	4.14	compound (42), periodic table (41), metal (31), atom (24), substance (16), mixture (14), gas (11), hydrogen (11), group (11), sodium (11), pure (11)
ATOMIC NUMBER	4.13	proton (55), atomic mass (40), periodic table (33), electron (33), neutron (32), atom (25), element (17), mass number (14), mass (12)
HALOGEN	4.04	chlorine (58), bromine (42), fluorine (33), iodine (32), group (29), gas (25), seven/7 (25), periodic table (19), hydrogen (11)
FORMULA	4.02	equation (56), symbol (24), chemical (21), valency (17), compound (17), element (13), number (13), balance (11), molecule (11)
REDUCTION	3.95	oxidation (61), oxygen (49), hydrogen (41), electron (40), lose/gain/donate (30), gain/accept (26), redox (12)
NEGATIVE	3.94	positive (86), ion (41), cathode (36), electron (35), charge (21), anode (13), electrolysis (13)
REDUCING AGENT	3.93	reduction (40), oxidising agent (36), hydrogen (33), oxygen (31), oxidation (27), electron (26), loss etc (21), gain etc (20), sulphuric acid (17)
ATOMIC MASS	3.71	proton (42), neutron (38), number (31), atom (26), electron (24), atomic number (24), periodic table (17), element (15), formula (14)
ISOTOPE	3.68	carbon (24), neutron (23), atom (19), electron (17), proton (17), chlorine (16), atomic number (12), element (12), radioactive/ity (11), diamond (10), [atomic mass (9)]
ELECTROCHEMICAL SERIES	3.55	reactivity (31), metal (23), sodium (19), element (16), potassium (14), electrolysis (13), hydrogen (11), electricity (10), periodic table (10), order (10)

In terms of the particular associative meanings of the different stimulus words a number of points of interest may be noted. While many words had a few "strong" associates (in Table 7.1 bold type has been used to highlight the responses used by more than half the students) a number had rather less sharply defined associative meanings. Thus for the following concepts there was no associate which was common to as many as half the O-level chemistry students taking part in the study: ISOTOPE, ELECTROCHEMICAL SERIES, VALENCY, REDUCING AGENT, ATOMIC MASS, ELEMENT, ELECTRIC CURRENT, ATOM. Most of these were words of relatively lower "meaningfulness" although the last two mentioned were not. At the other extreme it was of interest to note some of the strongest associative meanings. Although attempts at classifying associative responses have rarely proved satisfactory (see Section 5.2.5) two groups of strong responses must be noted. Thus words of opposite or reciprocal meaning were often elicited (eg POSITIVE - negative; ANODE - cathode; OXIDATION - reduction) and, more interestingly, words which defined or described the stimulus word were also common (eg CATHODE - electrolysis, negative; ATOMIC NUMBER - proton; PERIODIC TABLE - element, group. More general associates (eg PROTON - neutron) and particular examples (eg HALOGEN - chlorine) were less prominent as strong associates, although more frequent lower down response hierarchies.[1]

7.2.2 **Hierarchical cluster analyses.** In evaluating these and other cluster analysis methods on the basis of data relating to the test population as a whole, a readily interpretable map was sought on the assumption that idiosyncratic differences in the maps of individuals, or between particular teaching sets, would tend to be averaged out. Thus the population map could be expected, at least to a first approximation,

---

1. Table 7.1 may be loosely regarded as representing the response hierarchies of an "average" student.

to mirror the subject-matter structure as reflected in the diagrams constructed when planning the PS test (Figures C1 - C3 in Appendix C). Support for such an expectation was found in a number of the studies reviewed in Section 5.2 (eg Shavelson, 1972; Preece, 1976c; Fisher, 1979) in which group maps produced recognisable approximations to a priori subject-matter structures. Although the structures drawn in this case depended in detail on the subjective judgement of the author, they were felt to provide an acceptable base-line from which to work; and although they were neither simple nor fully hierarchical, the presence of hierarchical elements was held to justify a preliminary examination of hierarchical clustering methods. Amongst such techniques the nearest-neighbour method and McQuitty's group average method have been most widely used and are also to be preferred on theoretical grounds (see Section 5.2.7).

These two methods were therefore applied to a half-matrix of mean relatedness coefficiencts for the test population as a whole, using the CLUSTAN 1B suite of programs (Wishart, 1979). The resulting dendograms are shown in Figures 7.1 and 7.2 and are discussed below.

The nearest-neighbour method (Figure 7.1) groups concepts on the basis of their highest relatedness coefficients. Four recognisable clusters were immediately apparent on inspection of the dendogram: an electrolysis cluster consisting of nine concepts, an atomic structure cluster of five concepts, a periodic table cluster of four and a redox cluster of three. The two former clusters showed clearly interpretable hierarchical elements differing only in detail from that in the a priori structures. The logic of the hierarchy was less apparent in the periodic table and redox clusters. A characteristic of the nearest-neighbour method is its tendency to produce "chains" and six concepts

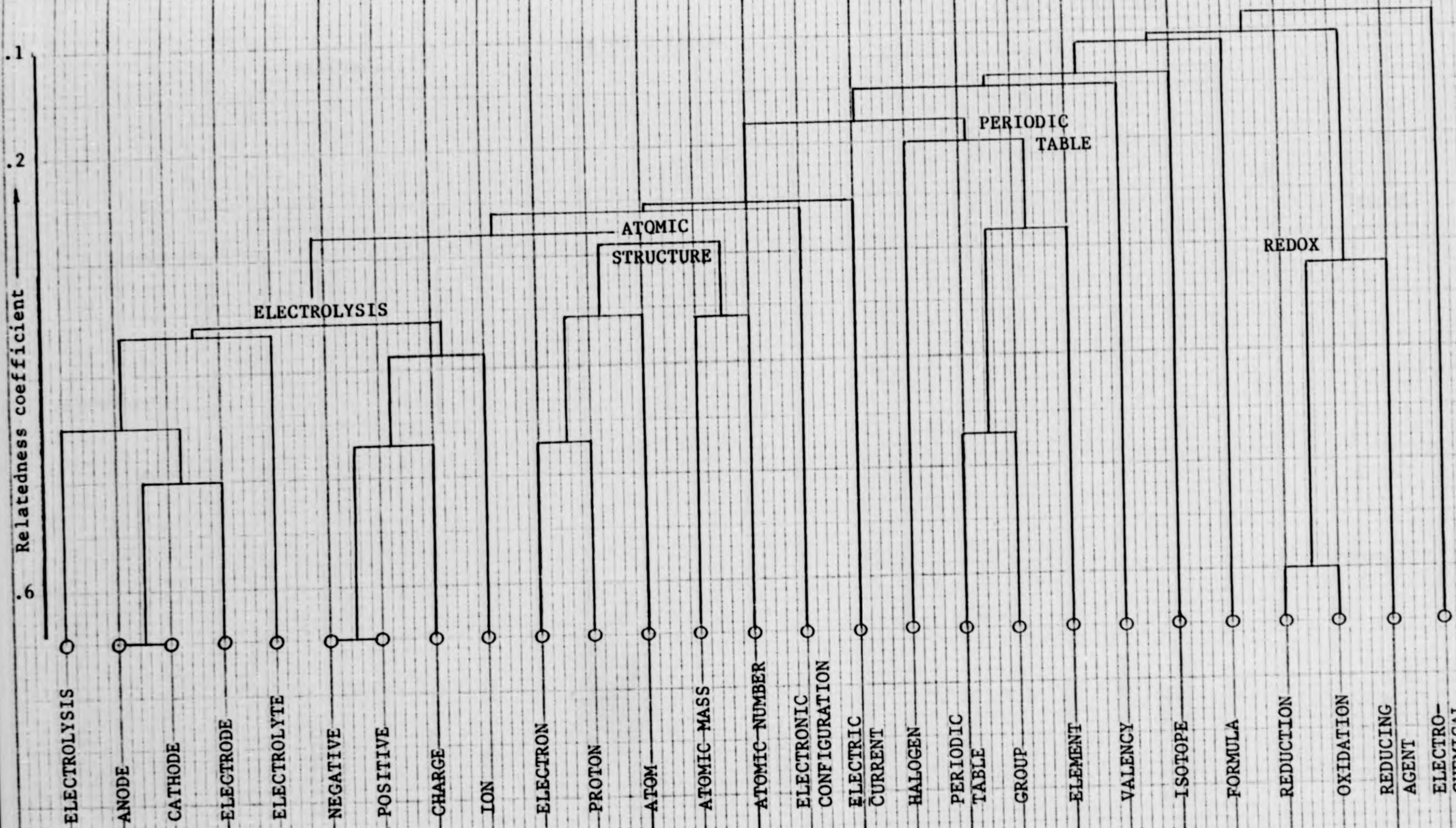


Figure 7.1 Cluster analysis of mean matrix; nearest-neighbour method

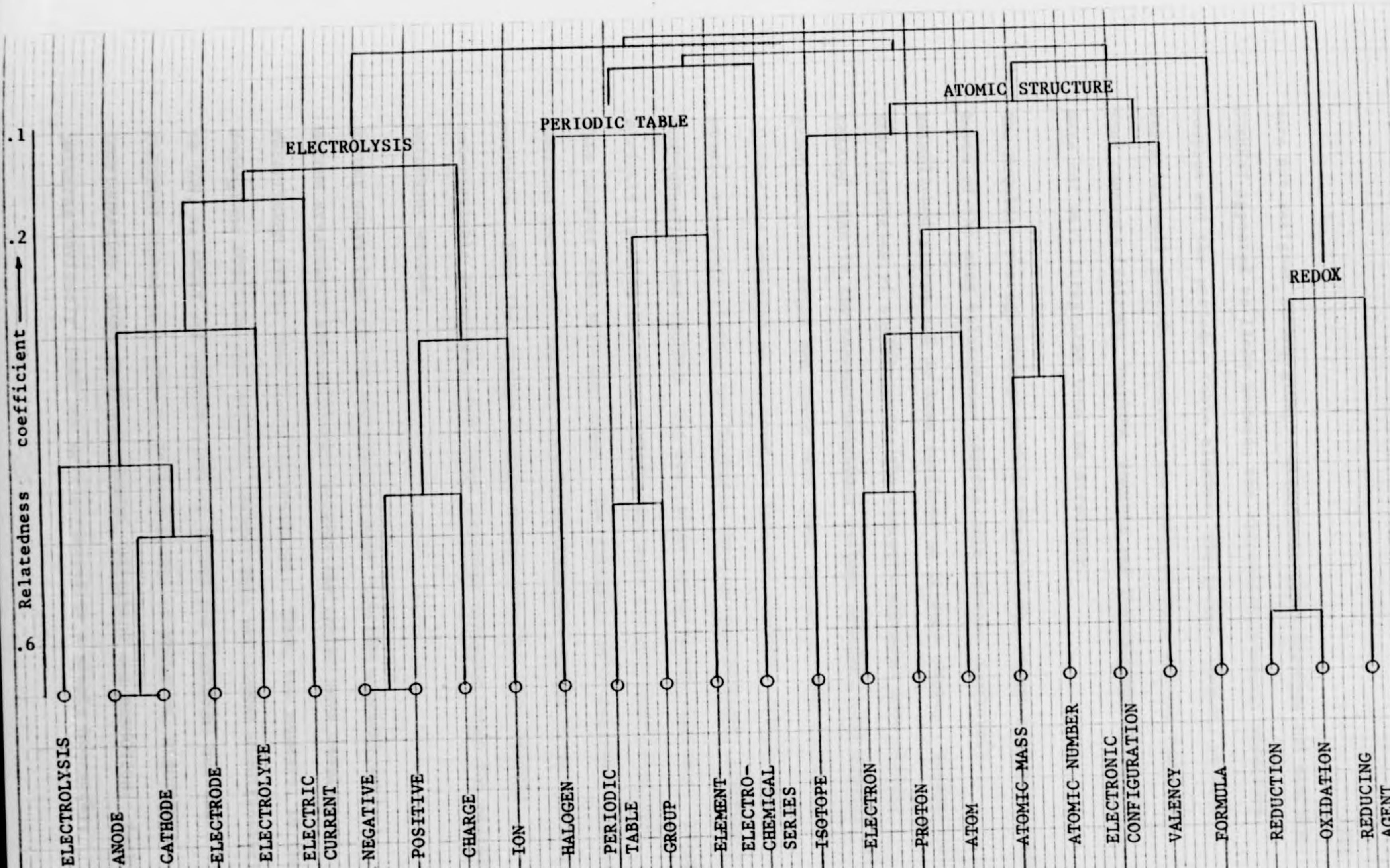


Figure 7.2 Cluster analysis of mean matrix; McQuitty's average linkage method

were not assigned to any group. ELECTRONIC CONFIGURATION and ELECTRIC CURRENT appeared as links between the electrolysis/atomic structure complex and the periodic table cluster, the former concept, at least, appearing well placed. However, the redox cluster and the remaining four concepts (VALENCY, ISOTOPE, FORMULA and ELECTROCHEMICAL SERIES) were not well integrated into the structure.

The McQuitty average linkage method (Figure 7.2), groups concepts on the basis of the **mean** relatedness coefficients between previously established groups and is less inclined to produce chains. Because of this averaging of coefficients, the concepts appeared less closely related than in the nearest-neighbour method, however, very similar groupings were found and there were fewer isolates. The electrolysis cluster was expanded to accommodate ELECTRIC CURRENT. Similarly, the atomic structure cluster now included the concepts ISOTOPE, ELECTRONIC CONFIGURATION and VALENCY, the two latter being paired together. Though strictly speaking isolated, FORMULA was more closely associated with this cluster than with any other. The periodic table and redox clusters remained unchanged, leaving only ELECTROCHEMICAL SERIES in an isolated position. Thus the average linkage method provided more readily interpretable groupings, with fewer isolated concepts, than the nearest-neighbour method.

Direct comparison of the associative maps with the a priori subject-matter structures were difficult because the latter were complex and interlinked rather than fully hierarchical. Thus, for example, in the dendograms POSITIVE and NEGATIVE CHARGES were not linked simultaneously to ION, to ANODE and CATHODE, and to PROTON and ELECTRON; nor was ELECTRON linked in a complex pattern to all of NEGATIVE, ELECTRONIC CONFIGURATION, ATOM, ELECTRIC CURRENT, OXIDATION and REDUCTION. Such patterns obviously could not be represented in a simple hierarchy, nevertheless the convergence of the maps and the subject-matter structure was impressive and the electrolysis, atomic structure and periodic table clusters in the average linkage map differed only slightly from the corresponding subject-matter structures. This provided strong evidence that the word associations of the students were reflecting the structure of the chemistry they had learned and encouraged the investigation of alternative, non-hierarchical cluster analysis solutions.

7.2.3 **Non-hierarchical cluster analyses.** Two non-hierarical clustering methods were discussed in Section 5.2.7. The first was a method described by Jardine and Sibson (1968) which identifies "maximal complete sub-groups" of entities which are fully interconnected above a series of given threshold levels and allows such groups to intersect at a predetermined number of points; the second, a stepwise graphical method described by Waern (1972) in which all pairs of entities interconnected above a series of threshold levels are connected by lines, further lines (to new entities only) being added as the threshold is lowered step by step. While Jardine and Sibson's method demands considerable computer time, Waern's method can be carried out quite rapidly by hand from a relatedness half-matrix. The two methods are, in fact, almost equivalent but the former produces a series of Venn diagrams while the latter produces a graphical solution. Trial runs with both quickly established that they produced virtually identical solutions and that the graphical version was the more readily interpreted. Particularly towards the end of an analysis, as lower thresholds became involved, Jardine and Sibson's method produced large, unstructured clusters in which the information available at earlier stages in the analysis was lost. In contrast, Waern's method preserved information from step to step. Thus only the results produced using Waern's method are discussed below.

Waern (1972) suggests that cut-off thresholds should be selected so as to (i) minimise the number of isolated entities, and (ii) maximise the distance between the two most distant points on the graph. Striking a balance between the two criteria resulted in a six step analysis using successive threshold values for the relatedness coefficients as listed below:[1]

Threshold	level	1	- relatedness	coefficient	> 0.40
"	"	2	"	"	> 0.30
"	"	3	"	"	> 0.25
"	"	4	"	"	> 0.20
"	"	5	"	"	> 0.15
"	"	6	"	"	> 0.10

These values were similar to those employed in other recent investigations (eg Preece, 1976a, b, c; Cachapuz, 1979; Fisher, 1979; Moynihan, 1982; Kempa and Nicholls, 1983). The resulting graphs, shown in Figure 7.3, are discussed below and comparisons are made with the a priori subject-matter structures produced during the planning of this phase of the study (Figures C1 - C3, Appendix C). While it is customary to label each line on the graph with the number of the step in the analysis at which it was added, it was found more convenient to label, at each step, only newly added lines.

---

1. The meaning of these threshold values of relatedness coefficients (RC) may be clarified by reference to the following descriptions. These are representative only and are based on response lists of average length.

   RC = 0.5 - each word is probably the first or second response to the other; if not then the response lists are near identical.

   RC = 0.3 - one word may be the first response to the other or both words may elicit the same first response etc.

   RC = 0.2 - the first response to one word may be the second response to the other.

   RC = 0.1 - the lists may have their third responses in common etc.

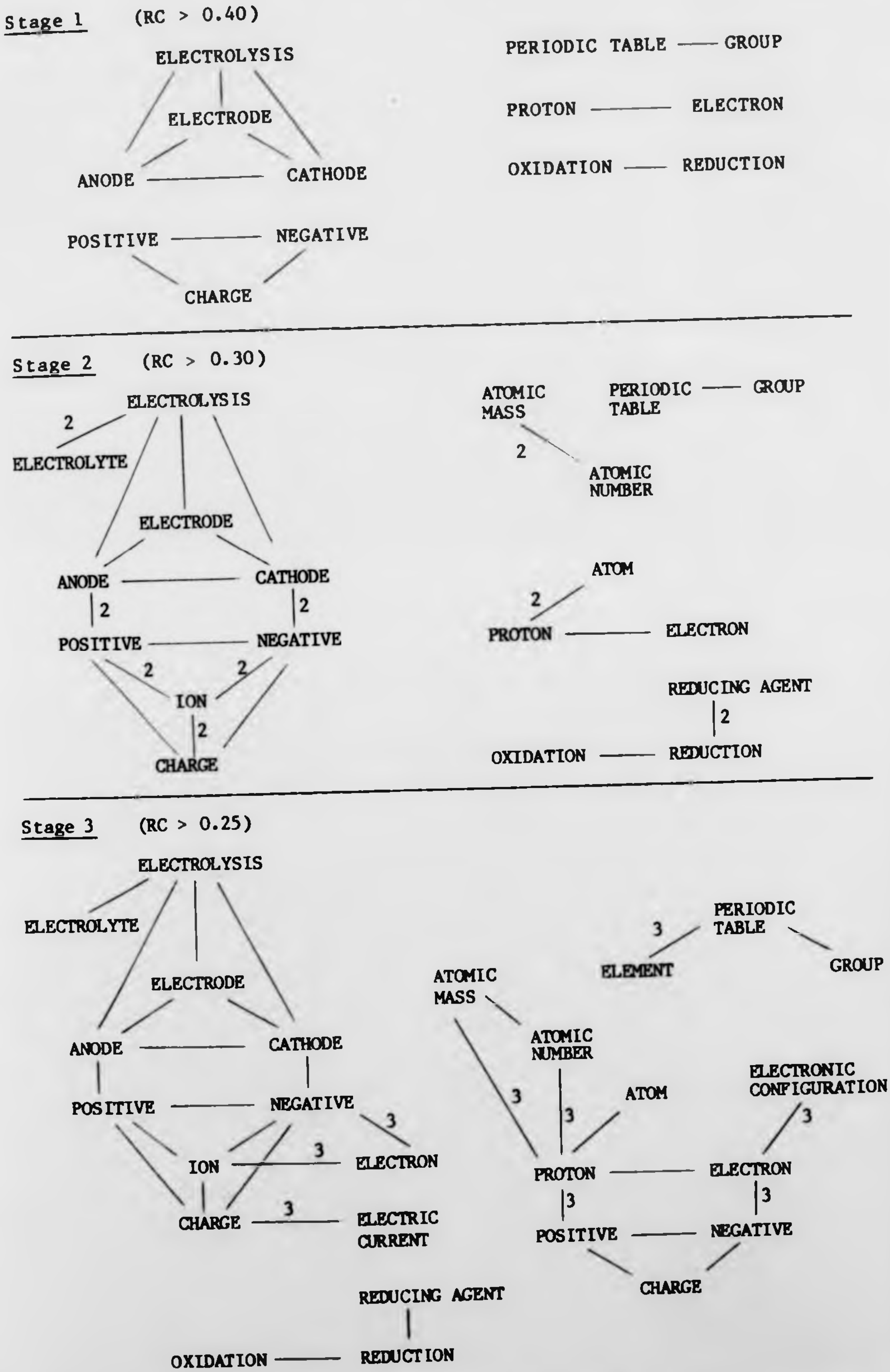
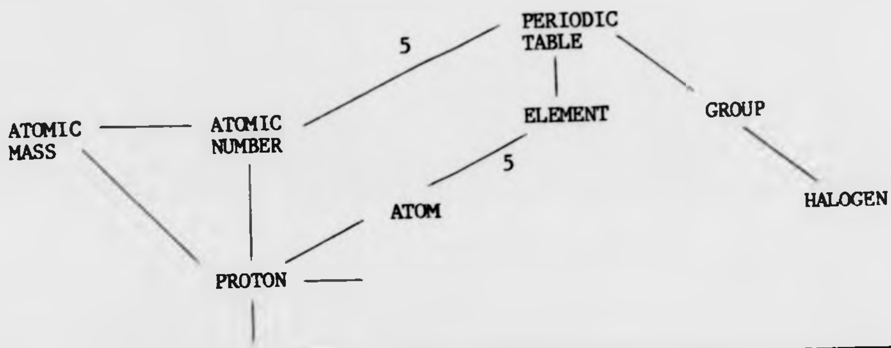


Figure 7.3 Cluster analysis of mean matrix; Waern's method

Stage 4 (RC > 0.20) - one new link only

Stage 5 (RC > 0.15) - two new links only



Stage 6 (RC > 0.10) - no new links to electrolysis cluster

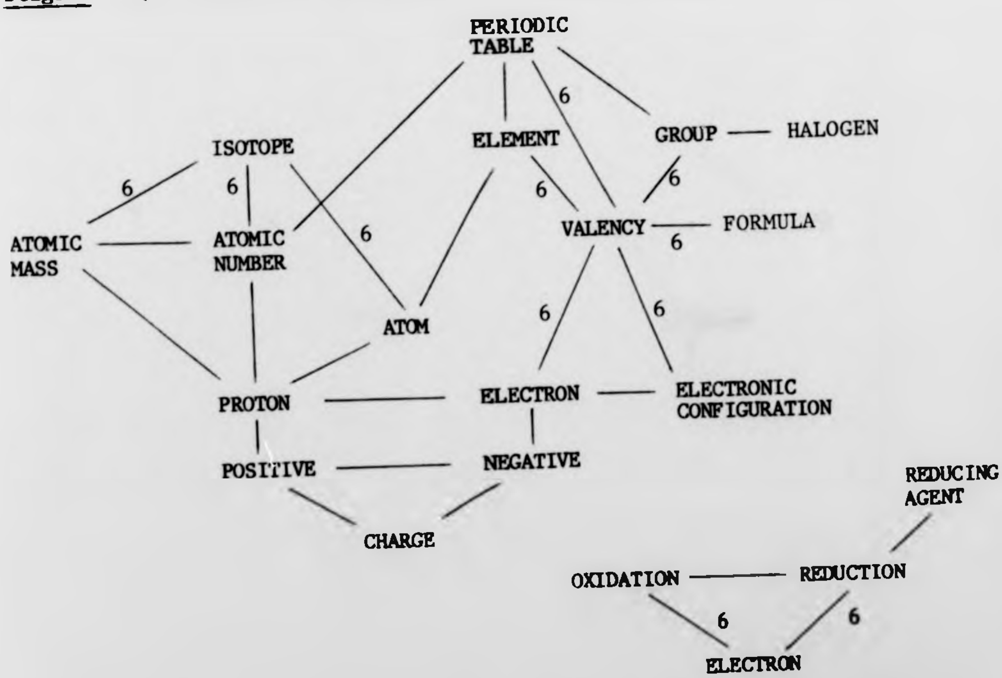


Figure 7.3 Cluster analysis of mean matrix; Waern's method (cont)

At the first stage of the analysis, five small clusters were observed. Of these an "electrolysis" cluster interlinking four concepts was particularly noted. At the second stage this cluster expanded, and by the third stage, beyond which no further elaboration occurred, showed close affinity with the a priori subject-matter structure. Of the concepts included in the word association test only ELECTROCHEMICAL SERIES was missing, and it may be conjectured that this concept was unfamiliar to most of the students involved. The associative structure showed a few additional links which had not been predicted. Apart from "opposites" (ANODE - CATHODE, POSITIVE - NEGATIVE), which are a well-known feature of word associations, these were readily interpretable in terms of the logic of elementary chemistry.

By the third stage in the analysis, three further clusters were becoming elaborated. Those containing concepts related to the periodic table and to redox systems are discussed later; the third clearly represented atomic structure. Insofar as four of the concepts involved in this cluster were shared in common with the electrolysis cluster, the clarity of the graph was improved by representing such concepts separately for the two clusters. It can be seen that the map of "atomic structure" closely resembles the corresponding a priori subject-matter structure, lacking only the suggested direct links between ATOM and ELECTRON, and between ATOMIC NUMBER AND ATOM. These links, in fact, only narrowly failed to appear at the second and third stages of the analysis respectively. Stage 4 of the analysis added HALOGEN to the "periodic table" cluster and stage 5 linked this cluster to the "atomic structure" group through links between PERIODIC TABLE and ATOMIC NUMBER, and between ELEMENT and ATOM, exactly as suggested in the a priori subject-matter structure. At stage 6, three important features were added.

First, the concept ISOTOPE joined the "atomic structure" cluster very much as expected. Second, the concept VALENCY provided additional links between the "periodic table" and "atomic structure" clusters although six links, rather than the suggested three, were involved. Finally, the "redox" cluster became linked to the concept ELECTRON. The latter has been separately represented in the map to preserve the clarity of the graph and to facilitate comparisons with the a priori subject-matter structure.

The convergence of the structures of the three interpenetrating clusters with the a priori subject-matter networks was very striking. If the associative links of reciprocally related concepts are ignored, the structures are near identical in many respects and all links in the associative map are readily interpretable in chemical terms. Given the assumptions already referred to regarding, (i) the acceptability of the a priori subject-matter structures as reflecting the interrelationships taught to O-level chemistry students, and (ii) the averaging effects of using the half-matrix of mean relatedness coefficients, this convergence strongly indicates the validity of the method employed as a means of mapping learned semantic structures. Waern's method obviously represented a significant improvement on hierarchical clustering methods in this respect and was therefore adodpted for the main analyses which are reported in Section 7.4 below.

7.3 Some non-structural comparisons of the word associations of high and low achievers

In this section, non-structural comparisons are made of the word associations of problem-relevant concepts for students of high and low achievement on these problems. Both "meaningfulness" and "associative meaning" are considered.

Groups of high and low achievers were identified with reference to the first four questions on the PS test. These comprised 21 items relating to the topics electrolysis, redox and the periodicity of atomic structure, and had been used as the basis for selecting concept words for the WAT. Of the items concerned, it has already been noted that item 1(c) (104) was unclear to many students and that items 4(c)(i), (ii) and (iii) (404-406) did not discriminate well.[1] The distribution of scores on the remaining 17 items was computed (mean 8.37, standard deviation 3.33) and high and low achievers were selected on the basis of scores exceeding one standard deviation above and below the mean, respectively. The "top" group comprised 26 students with scores of 13 or more out of 17 (mean 14.1, standard deviation 0.99) and the "bottom" group 29 students with scores of 4 or less (mean 3.14, standard deviation 1.03).[2] The difference between the mean scores of these groups was highly significant statistically ($t = 40.23$; $df = 54$; $p < 0.001$).

Although this procedure distinguished very satisfactorily between high and low achievers on the problems as a set, it was felt that discrimination might be further improved by selecting separate groups on the basis of performance on each of the three topics. This was done, for each topic, on the basis of scores on the five items whose solution was judged to be most closely dependent on the concepts included in the WAT. Groups of between 18 and 23 students were selected such that those in

---

1. Discussion of these items will be found in Section 6.1.1 and in Appendix B.

2. The four versions of the test differed little in overall difficulty, and for only a small minority of items did variations of the task formulation lead to significant differences in facility. As was reported earlier (Section 6.1.4) statistics calculated across test versions agreed closely with those calculated separately for each version. The selection of high and low achievement groups without reference to test versions was therefore considered justifiable.

the "top" groups all had perfect scores on their set of five items, while those in the "bottom" group all had zero (or, in the case of the "electrolysis" set, either zero or one).[1] In practice, analyses based on these groups yielded results which were closely similar to those obtained using the original groupings. To avoid difficulties associated with concepts common to more than one topic area, and to keep the account reasonably concise, the results reported below are based largely on the original groupings. However, reference is occasionally made to data from the separate topic groups where this adds to the picture.

7.3.1 "Meaningfulness" and mean relatedness coefficients. Table 7.2 shows the mean lengths of the response lists and the mean relatedness coefficients for the high and low achievement groups, together with their scores on the relevant set of test items. While high achievers appeared to have somewehat longer response lists, that is to say, more "meaningful" concepts (Noble, 1952), this effect did not quite reach statistical significance at the 5% level and must be regarded with some caution. There was no difference in the mean overlap between concepts as reflected by the relatedness coefficient. The latter finding is consistent with that reported by Rothkopf and Thurner (1970), who found no correlation between PS scores and mean associative overlap. Both findings are also consistent with Preece's (1977) interpretation of the development trends of certain indices derived from word association tests (see Section 5.2.5). He implies, in effect, that neither "meaningfulness", nor mean associative overlap, can distinguish between

---

1. It may be of interest to note in passing that two students appeared in the top group (perfect score) in one topic area and the bottom group (zero score) in another; a further ten students in one or other bottom group, averaged 70% or more on the other two topics. This further confirms evidence reported earlier regarding the topic-specific nature of examination performance.

well and poorly discriminated associations. The question of the **quality** of the associations of students of high and low achievement is taken up in the next section.

Table 7.2 Quantitative data relating to the word associations and PS test scores of students of high and low achievement

	Top group		Bottom group			
	Mean	SD	Mean	SD	t	df
Mean length of response list	5.12	1.35	4.53	1.10	1.77+	53
Mean relatedness coefficient	.086	.018	.084	.027	0.39	53
Score on relevant items (ex 17)	14.12	0.99	3.14	1.03	40.23*	53

* $p < 0.001$; + $0.05 < p < 0.10$

7.3.2 **Associative meaning.** The WAT responses of the high and low achievement groups reflected a number of significant differences in the associative meaning which each group attached to most of the stimulus words. Figure C5 in Appendix C shows, for each group, the frequency (%) of all responses used by at least 30% of either. Where this allowed less than five responses to be shown, the threshold was lowered to 20%. Two shortened examples (using a threshold of 40% and based on data from the separate topic groups) are shown in Figure 7.4 for illustrative purposes. Each stimulus word is printed in capital letters, and is followed, in lower case, by responses which it elicited. These are ranked according to the frequency with which they were used by the top group. The proportion of each group using each response is indicated by a bar graph. The significance of the difference between the proportions of high and low achievers making any particular response was tested using a method involving the calculation of the standard error of the differences between proportions (Lewis, 1973, p 122).

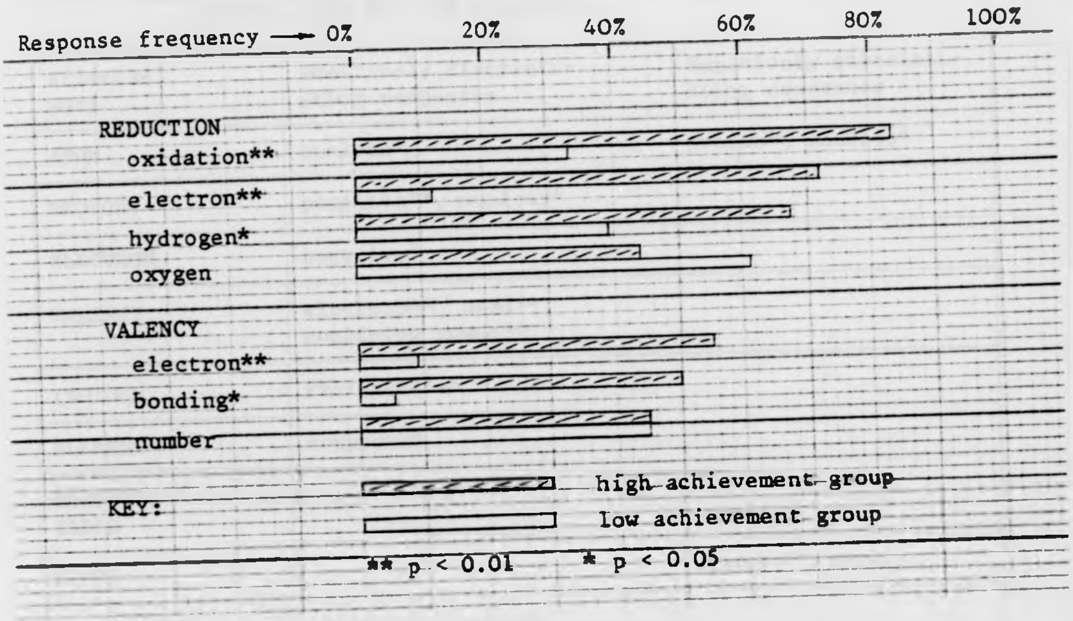


Figure 7.4 Most frequent responses (%) of high and low achievement groups (separate topics) to two stimulus words

It can be seen from Figure 7.4 that a number of the responses shown were used significantly more often by high achievers than by low achievers. These include "oxidation", "electron" and "hydrogen" as responses to REDUCTION, and "electron" and "bonding" as responses to VALENCY. The response "oxygen" to REDUCTION shows the opposite trend while the response "number" to VALENCY fails to discriminate at all between the two groups. Responses of the first and second kinds were described as positively and negatively discriminating respectively. Table 7.3 summarises all responses which discriminated between high and low achievers at statistically significant levels. A number of trends regarding differences between the associative responses of high and low achievers are briefly discussed below, with reference to this table, for the light which they throw on the mapping exercise reported in the next section. Further examples from Figure C5 which, though failing to reach

Table 7.3 Associative responses which discriminated significantly between high and low achievers

Stimulus word	Positively discriminating responses	Negatively discriminating responses
ATOM	nucleus*	proton*, neutron*
PROTON	positive**, nucleus*	
ELECTRON	negative*	
VALENCY	electron**, bond**, combining power**	
ISOTOPE	radioactive*	
ELEMENT	compound*	
PERIODIC TABLE		metal*
GROUP		element*, metal*
HALOGEN	chlorine*, iodine*	
ELECTROLYSIS	ion*	positive*
ELECTROLYTE	solution*	
ANODE	electrode*	
ION	charge*	
ELECTROCHEMICAL SERIES	sodium**	
NEGATIVE	ion**	
CHARGE	ion*	
OXIDATION	reduction*, electron**	
REDUCTION	oxidation*, electron*, hydrogen*, redox*	
REDUCING AGENT	electron**, oxidising agent*	

KEY: Significance levels * $p < 0.05$; ** $p < 0.01$

statistically significant levels, provided supporting evidence are also referred to.

The positively discriminating responses which characterised high achievers tended to be either:

(i) words directly used in defining or describing the stimulus (or which are defined or described by the stimulus);

(ii) important examples, sub-sets etc of the stimulus;

(iii) certain reciprocal associations;

(iv) other problem-relevant associations.

By contrast, the negatively discriminating responses which characterised low achievers tended to be poorly discriminated general associations.

Examples of the first class of positively discriminating responses were numerous. Thus "proton" is shown in the a priori subject-matter structure (Figure C2, Appendix C) as (partly) defined or described by reference to "nucleus" and to "positive". Both responses were significant positive discriminators while the vaguer associations "neutron" and "electron" tended to discriminate negatively (though not at a statistically significant level). Similarly "electron" was a much more likely response to both OXIDATION and REDUCTION for a high achiever than for a low achiever. A reciprocal example, in which a stimulus defined or described a positively discriminating response, was provided by the response "ion" to CHARGE. The prevalence of "directly related" responses (Preece, 1977) in the associations of high as opposed to low achievers was confirmed statistically. For this purpose "directly related" responses were defined as those joined to the stimulus word by a direct link in the a priori subject-matter structures (Figures C1 - C3 in Appendix C). The mean number of such responses for the high achievement group was significantly greater than that for the low achievement group.[1]

In the second class of positively discriminating responses were those which were important examples of the stimulus word. Thus high achievers were significantly more likely to respond "chlorine" or "iodine" to HALOGEN and "radioactive" to ISOTOPE.

Responses which can be regarded as opposites to the stimuli which elicit them are well-known in WATs. In the present test, most such responses tended to be weakly negative discriminators. This applied to the pairs "anode/cathode", "positive/negative" and "proton/electron". However, "oxidation" and "reduction" were positively discriminating mutual responses. By way of rationalisation it could be argued that oxidation and reduction are mutually interdependent concepts and that their association is therefore fundamental in a way which does not apply to the other pairs of concepts.

The fourth class of positively discriminating responses was described as "other problem relevant associations". To a large extent this covered both the previous classes and tended to overlap particularly with the first, although the definitions or descriptions involved were less clear cut. Thus the associations of "solution" with ELECTROLYTE, and "electron" with VALENCY, fall into this class.

Many responses discriminated little between high and low achievers, but a number were noted which tended to discriminate negatively, some at statistically significant levels. Such responses appeared to have few obvious characteristics in common and seemed best regarded as general,

---

1. High achievement group mean 1.01, standard deviation 0.60; low achievement group mean 0.74, standard deviation 0.51; $t = 6.69$, $df = 26$, $p < 0.001$.

poorly discriminated associations. The responses "metal" to PERIODIC TABLE and "positive" to ELECTROLYSIS were good examples. It was interesting that while the response "nucleus" discriminated positively as a response to ATOM, both "proton" and "neutron" were significant negative discriminators. This presumably reflected the fact that while the "nucleus" is part of an ATOM, "neutrons" and "protons" are parts of a NUCLEUS first, and thus only at a second order, of an ATOM.

The findings reported above appeared to augur well for the cognitive mapping exercise. Thus it seemed likely that the "directly related" responses of the high achievers would tend to lead to well discriminated associative overlaps and thence to maps closely reflecting the logical organisation of the stimulus concepts. Conversely the vague, poorly discriminated associations of the low achievers might be expected to introduce a random element into the associative overlap between concepts, yielding poorly discriminated structures. An account of the maps actually found for high and low achievers follows in the next section.

7.4 The associative maps of high and low achievers

In planning this phase of the study, it was hypothesised that the organisation of chemical knowledge represented by the cognitive map of the high achiever would differ from that of the low achiever in ways which could be related to their differing PS performance; more specifically that the maps of high achievers would be more systematic, with higher overall levels of linkage and superior discrimination between more and less closely related concepts (Section 5.4). These expectations were broadly confirmed and are discussed and elaborated below with reference first to the mean associative maps of the high and low achievement groups and subsequently to the maps of individual high and low achievers. An attempt is also made to explore particular PS episodes in the light of individual cognitive maps.

7.4.1 Descriptions of the mean maps of high and low achievement groups.

Mean associative maps for the high and low achievement groups were generated using Waern's stepwise method as described earlier in the chapter. They are presented in Figures 7.5 and 7.6 (pp 288 and 289). To facilitate comparisons, the structures emerging after the second, fourth and sixth stages of the analysis have been shown.

The maps of the two achievement groups showed broad structural similarities; however, two important differences were apparent. At any given level of relatedness between concepts, the low achievement group tended to have less interconnected structures than the high achievement group, and important linkages between "directly related" concepts were often missing. In this context "directly related" concepts were defined by the a priori subject matter-structures, Figures C1 to C3, in Appendix C and were concepts judged to be of primary importance in defining or describing one another in the context of a GCE O-level chemistry course. Secondly, a number of concepts were linked rather differently in the two structures, that of the low achievement group tending to contain more links between concepts which were not "directly related". A general discussion of the ways in which the first of these differences can arise, in the context of Waern's method of analysing a relatedness matrix, must precede a detailed and specific examination of the two structures.

Waern's method demands that links be drawn between concepts which are interrelated above a series of decreasing threshold levels, providing only that the concepts concerned have not already been linked into a common structure at an earlier theshold. Thus, the absence of a link at a given stage of the analysis may simply indicate that the concepts concerned are related (if at all) only below the particular threshold level chosen. Following a number of precedents, all analyses in the present study were terminated at a relatedness coefficient of 0.1. Thus any links represented by lower coefficients would be too weak to appear. However, the persistent failure of a particular link (say of X to Y) to appear at any step during an analysis might also have a quite different explanation; namely that their interrelationship was not a prominent one relative to their separate interrelationships with other concepts. Thus if both X and Y have other much stronger associates and have been linked through these into a common structure at a higher threshold, then the **relatively** weaker X-Y link will not appear.[1] Insofar as the accessibility of one idea from another was considered likely to influence PS behaviour, the tendency of the method of analysis employed to suppress associations that were weak, either in absolute terms or in relation to competing associations, was considered not merely acceptable but, indeed, appropriate.

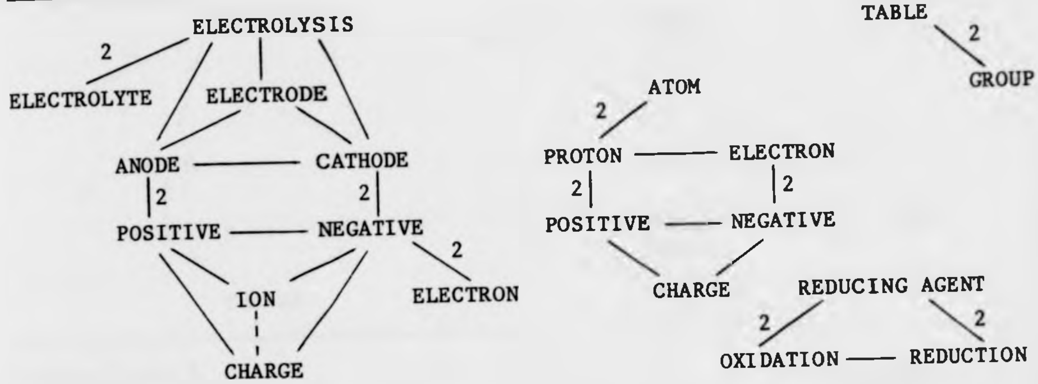
The associative maps of the high and low achievement groups (Figures 7.5 and 7.6 respectively) can now be discussed in detail. It will be convenient to compare them topic by topic.

(i) **The electrolysis cluster.** After the first two steps of the analysis, the maps of both achievement groups showed a pattern of concepts similar to that obtained for the population as a whole (Section 7.2.3). However, for the high achievement group some of the links were stronger

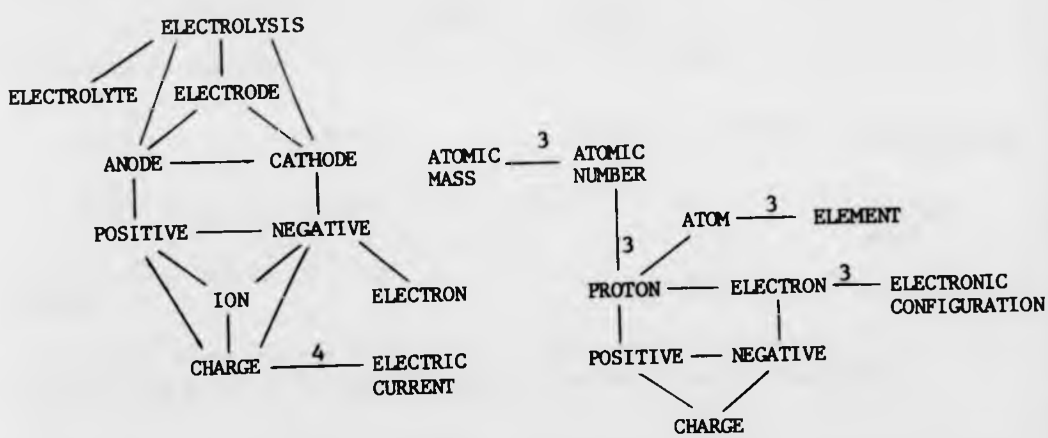
---

1. There was obviously a danger that the application of a particular threshold might distort an analysis by failing to record one amongst a number of associations of similar strength, simply because it happened to fall marginally below the nominated level. Cases of this kind were, in fact, easily detected when constructing the maps and were fortunately rare. Where they occurred, and providing that neither of the concepts concerned was involved in any other interrelationships at above the next lower threshold level, it was deemed more useful to add the link to the map (using a dotted line) rather than to allow the relationship to be obscured. One such example appears in each of Figures 7.5 and 7.6.

Stages 1 and 2



Stages 3 and 4 - "periodic table" and "redox" clusters unchanged



Stages 5 and 6 - "electrolysis" cluster unchanged

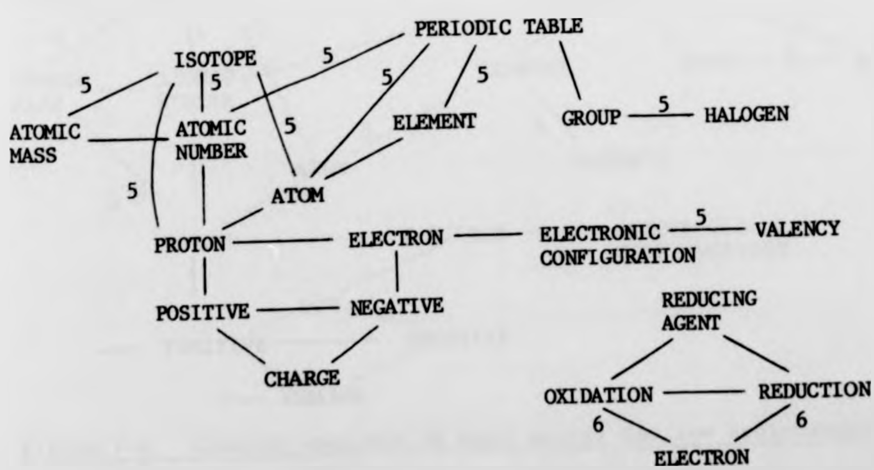


Figure 7.5 Cluster analysis of mean matrix for high achievement group

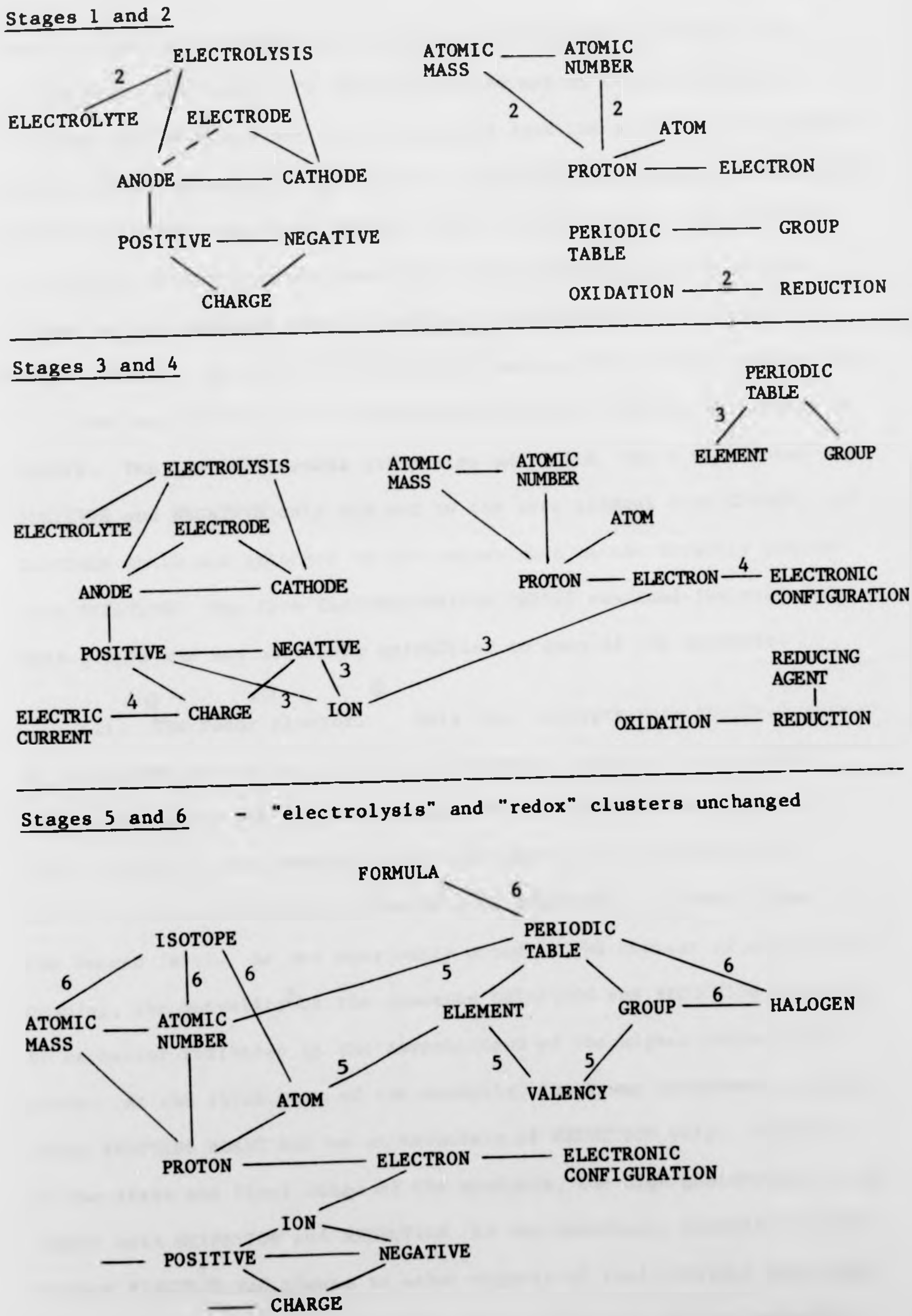


Figure 7.6 Cluster analysis of mean matrix for low achievement group

and the concept ELECTRON was already attached to the structure. By comparison, six links were absent from the map of the bottom group. The concept ELECTRODE was not well integrated into the structure and neither ION nor ELECTRON were present at all; CATHODE was not linked with NEGATIVE (in fact it was associated equally with both NEGATIVE and POSITIVE at the second threshold level, but these undiscriminating links did not appear as the concepts were already part of the same structure at the higher level). At the third and fourth levels, beyond which neither map developed any further, both groups added ELECTRIC CURRENT by linkage to CHARGE. The low achievement group also added ION, which was linked to POSITIVE and NEGATIVE only and not to the more general term CHARGE, and ELECTRON which was attached to ION rather than to the directly related term NEGATIVE. The term ELECTROCHEMICAL SERIES remained isolated for both groups and was obviously unfamiliar to many of the students.

(ii) **The redox cluster.** Only four concepts were involved here. By the second threshold, the high achievement group had established a simple triangular map in which OXIDATION and REDUCTION were linked at the first level, and REDUCING AGENT with both at the second level. For the low achievement group, OXIDATION and REDUCTION only were linked at the second level. As was previously noted in the context of associative meaning, the mutuality of the concepts OXIDATION and REDUCTION appeared to be better reflected in the associations of the higher achievement group. At the third step of the analysis, the lower achievement group added REDUCING AGENT but as an associate of REDUCTION only. Finally, at the sixth and final stage of the analysis, the high achievement group linked both OXIDATION and REDUCTION to the important, directly related concept ELECTRON and thence to other aspects of their overall knowledge structure. For the low achievement group the redox cluster remained a simple isolated chain of apparently very limited use in the context of PS.

(iii) **The atomic structure/periodic table cluster.** This was a more elaborate cluster in which up to 16 concepts were involved. As with both the other clusters, the map of the high achievement group, after the second stage of the analysis, was similar to, but rather further advanced than, that of the population as a whole. In particular, the atomic structure cluster included the correct charges linked to the sub-atomic particles concerned. These charges were not present in the map of the low achievement group which did, however, contain a unique triangular structure involving ATOMIC NUMBER, ATOMIC MASS and PROTON. It was tempting to regard the ATOMIC NUMBER - ATOMIC MASS link as resulting from crude general associations to the word ATOMIC and this was checked by reference to the response frequency graphs for the two achievement groups (Table C5, Appendix C). These showed that the high overlap was due, in roughly equal measure, to negatively discriminating mutual associations of each term for the other, and to their common associations with "proton" and "neutron", three out of the four observed pairings of response to stimulus reflecting second order, rather than direct, relationships and being negatively discriminating. Both low and high achievement groups had a second simple cluster at this level, consisting of PERIODIC TABLE and GROUP.

At the level of the third threshold, the structure of the high achievement group included ELECTRONIC CONFIGURATION linked to ELECTRON, ELEMENT linked to ATOM, and ATOMIC MASS and ATOMIC NUMBER linked to one another, and through the latter only, to PROTON. The PERIODIC TABLE cluster was unchanged. For the low achievement group ELECTRONIC CONFIGURATION was linked to the map in the same way, though only at the fourth threshold level. ELEMENT was linked to PERIODIC TABLE instead of to ATOM, however, as it is directly related to both, this was probably of little significance and in fact made no difference to the structure which finally emerged. Also at the third level, ELECTRON became linked to ION and thence to both POSITIVE and NEGATIVE (in the absence of this indirect link, ELECTRON would have linked directly to both POSITIVE and NEGATIVE at the fifth level!).

At the fifth and sixth threshold levels, four important groups of linkages occurred. VALENCY became linked at step five, to ELECTRONIC CONFIGURATION in the structure of the high achievement group, but less meaningfully to both ELEMENT and GROUP in that of the low achievement group. HALOGEN became linked to GROUP in the high achievers' map at the fifth threshold level and, less discriminatingly, to both GROUP and PERIODIC TABLE at the sixth level in the low achievers' map. ISOTOPE became linked to ATOMIC NUMBER, ATOMIC MASS, ATOM and PROTON at the fifth level in the map of the high achievement group and to all but PROTON at the sixth level in the low achievers' map. In both the latter cases it was noted that the links were weaker for the low achievement group. It may also be relevant that the protocols collected during the first phase of the study suggested that many students appeared to define isotopes by direct reference to numbers of protons and neutrons. If this was the case, then proton would be a directly related response for such students. The periodic table and atomic cluster structures were combined at the fifth level for both groups in essentially the same way. Finally FORMULA, which remained isolated for the high achievement group, was linked rather undiscriminatingly to PERIODIC TABLE at the sixth level in the map of the low achievement group.

Following the rather detailed account above, the next section is concerned with summarising and interpreting these findings.

7.4.2 An interpretation of findings in relation to the mean maps. The mean associative map of the high achievement group differed from that of the low achievement group in that it tended to have more and stronger links between directly related concepts and fewer links between concepts which were not directly related. A visual comparison of the maps of the two groups showed that 19 links between directly related concepts were weaker or absent in the map of the low achievement group as compared to four which were weaker in that of the high achievement group.[1] The map of the low achievement group also contained four links, absent in that of the high achievers, between concepts which were not directly related. As a result, the map of the low achievement group was less complex and more fragmented than that of the high achievement group. Comparisons of the associative meanings of the concepts for the two groups (Section 7.3.2) suggested that linkages involving important subsets and examples of concepts would probably have discriminated between high and low achievers in a similar way had it been possible to include some of the these in the WAT.

An assumption underlying this phase of the study was that the overlap of a student's lists of associative responses to two concept words (weighted to allow for the primacy of the responses) provided a useful estimate of the accessibility of one concept from the other in the student's LTM store; also that the inter-accessibility of particular problem-relevant concepts was a necessary, if not a sufficient, condition for succesful PS.[2] In terms of these assumptions it was

---

1. Of these four, three were nevertheless still strong links (RC > 0.25), and the fourth reflected the fact that the high achievers associated ELEMENT more closely with ATOM than with PERIODIC TABLE, arguably preferring the more fundamental relationship.

2. It could not be a sufficient condition insofar as (i) the nature of the recalled relationship might be inimical to success and (ii) errors might be introduced at other stages in the PS process.

apparent that the mean cognitive map of the high achievement group was better adapted for solving chemical problems in the topic areas concerned than was that of the low achievement group. To this extent the assumptions appeared to be vindicated. These points are illustrated below with reference to three examples, chosen almost at random, from the PS test.

Item 2(b)(202) asked students to identify a pair of isotopes on the basis of data relating to the number of electrons in a neutral atom and to mass numbers. The maps showed that high achievers had a more strongly defined concept of isotope and were more likely to associate isotope directly with proton (and, on the basis of data from the response frequency graphs in Figure C5, Appendix C, with neutron). It can also be seen from the maps that they were much more immediately aware of the charges on the proton and electron and were thus more likely to equate the numbers of each present in a neutral atom. The introduction of irrelevant formulae in place of mass data in the augmented (irrelevant) version of this item depressed the scores of the low achievers much more than those of the high achievers. They may have been distracted by what was obviously a poorly discriminated association of FORMULA with PERIODIC TABLE and attempted fruitlessly to make use of this data. The high achievers' map showed that FORMULA had no strong associations with any of the concepts in these topic areas and they may thus have been better able to ignore the irrelevant data.

Item 3(a)(301) asked for a definition of oxidation in terms of changes at the atomic level. It was immediately apparent that this would discriminate against those students whose cognitive maps contained a redox cluster which lacked ELECTRON and which was essentially isolated from the network relating to atomic structure. This, in fact, applied to all the redox items in the test, although obviously other factors would also be involved.

Finally item 2(e) (205) demanded the derivation of the formula for a compound on the basis of the numbers of electrons in the neutral atoms of two elements. Most of the students taking part in the first phase of the study who were successful in solving this item, made use of the data relating to electrons to derive the electronic configurations and thence the valency for each element. This route to the solution appeared much less likely for students in the low achievement group as the idea of the VALENCY was not directly accessible from ELECTRON and ELECTRONIC CONFIGURATION in the group map.

A difficulty in attempting to relate detailed PS behaviour to the cognitive maps presented so far concerns the averaging involved in generating the maps. Within a group of similar overall performance, different students tend to get their marks on different sets of items, and even where they succeed and fail on the same items, they may do so in different ways. A mean map for the population (Figure 7.3) may, for example, give the impression that "average" students have a rather rich associative network surrounding the concept VALENCY, involving potentially useful linkages of about equal strength to six other concepts. If so, this could be a misleading impression, a typical student having only two or three rather strong links and these links differing from student to student. For so complex an entity as a network of inter-connected associations amongst 27 concepts, an average map is only likely to resemble that of any particular individual if the variations between individual networks are fairly limited. Inspection of the associative maps of individual students suggested that they were often, by contrast, highly idiosyncratic in detail. Although the reliability of the word associations of individuals has not been established and a systematic study of individual maps would be beyond the scope of the present study, those of a small number of students are discussed in relation to their PS behaviour in the next section.

7.4.3 Some individual case studies. The trends demonstrated in the "average" associative maps of groups of students provided potential explanations for differences in PS performance which could be explored for specific items, however more detailed information was available from **individual** maps. The relationship between the PS behaviour of particular high and low achievers and their associative maps was therefore examined in a number of short case studies. To maximise discrimination between the two achievement levels, individual students were selected on the basis of either perfect or near-zero scores on all the items relating to a given topic. Information regarding the detailed PS behaviour of each was sought from their answer scripts and from their notes on "working" and "difficulties" on the accompanying self-reporting forms. As far as possible cases were selected for further study only when the information available from these sources was considered to give an adequate picture of the PS episodes, although this was only partially achieved.

As would be expected, the individual maps of high and low achievers showed similar features to those found in the corresponding "average" maps. In general, however, links between concepts tended to be both stronger and fewer. Although the maps of high achievers generally bore a closer resemblance to the average maps than was the case for low achievers, they were often less complex and less well structured than the average maps and showed considerable individual differences. However, the most prominent links were generally those linking directly related concepts. Thus, while a number of links to be found in the average map for high achievers might be absent, those which were present usually joined concepts standing in important defining or describing relationships to one another. The maps of individual low achievers, by contrast, tended to include idiosyncratic links, while links between directly related concepts were more likely to be weak or absent than was the case for high achievers. The maps of low achievers thus tended both to vary more widely and to be more fragmented than those of high achievers.

The individual maps often appeared to provide useful insights into particular PS episodes. This can best be illustrated with reference to examples and these have been provided in the ten short case studies which follow. Before discussing these, however, two limitations on the interpretive power of individual maps should be noted. The first concerns the more restricted nature of the inference which can be drawn from the presence of links in the associative map of an individual as opposed to that of a group. In both cases the inference is, strictly speaking, limited to the notion that one concept is more or less accessible from another. However, Nagy (1983) has argued that it seems implausible to suggest that word associations should successfully map subject-matter structures if such associations reflect false relationships. This argument has considerable force when applied to average maps, but is less persuasive in relation to individual maps which, even for high achievers, tend to contain idiosyncratic features and to mirror subject-matter structure somewhat less well. It has already been noted that the reliability of the word associations of individuals, as opposed to groups, has not been established. To this extent, more caution should probably be exercised when interpreting the maps of individual students.

A second restriction on the interpretive power of the maps of individuals, concerns the limited relevance of the maps to certain PS attempts. While it was often possible to account for "omitting" or "guessing" by reference to the absence of certain necessary associations (and even, in a few cases, to offer speculative interpretations for a particular "guess") other strategies sometimes by-passed the map available. In order to include a sufficient number of items and to cover a range of topics, only a few concepts not explicitly named in the items could be included in the WAT. In practice, a number of attempts depended critically upon relationships involving concepts which did not appear in the maps and this prevented a full interpretation of some PS episodes. The absence of data relating to the associations of NUCLEUS and NEUTRON was a case in point. It was clear that these concepts were relevant to a number of PS attempts for at least one item, but uncertain how their absence from a particular map may have distorted its structure.

In generating associative maps from relatedness matrices during the case studies, the cut-off thresholds employed were slightly modified as compared with those used with the average maps. Because of the generally higher coefficients involved in the matrices of individuals, the first threshold was raised to 0.50. It was found convenient to set four subsequent thresholds (identified as levels 2 to 5) at decrements of 0.10. Before turning to examples of the case studies, it may be helpful to amplify slightly the footnote on page 273 regarding the interpretation of each of these levels. A relatedness coefficient of 0.50, representing the first threshold, would be difficult to generate unless each of the pair of concept words involved, elicited the other as its first or second associative response. This expectation, based on calculations assuming response lists of various representative lengths, was confirmed by an inspection of the raw data from the WATs. Thus concepts linked at this level could be regarded as more or less immediately accessible from one another. At the second threshold (relatedness coefficient 0.40), one of the concepts tended to elicit the other, with both also sharing common associates of high primacy. For example, although ISOTOPE tended to elicit "atomic mass", ATOMIC MASS elicited "isotope" less often; both, however, tended to elicit "nucleus" as an early response. Thus, linkages at this level generally represented immediate accessibility in one direction (indeterminate from the matrix), and ready accessibility, though of a "second-order" through a third concept, in the other direction. The third threshold, at a relatedness coefficient of 0.30, was most often associated with concepts eliciting common first responses together with some further associative overlap. Links at this level were thought of as representing ready, mutual accessibility at a second-order level, that is, through a third concept. Links at levels 4 and 5, (relatedness coefficients 0.20 and 0.10) corresponded approximately to second and third associates, respectively, in common, plus some additional associative overlap in each case. These levels of linkage thus represented progressively weaker second-order mutual accessibility.

The organisation and presentation of the ten short case studies which follow, and which illustrate the value (and limitations) of individual associative maps in the interpretation of particular PS episodes, will now be explained. In Figure 7.7 (p 301) are reproduced the two sets of items involved.[1] These relate, respectively, to the topics ELECTROLYSIS and ATOMIC STRUCTURE/PERIODICITY. Each set also includes one item

1. Only the "standard" versions of the items are reproduced; where, occasionally, PS episodes related to "reduced" or "augmented" versions this is indicated (although evidence from word associations of any influence attributable to such variations was rare).

overlapping with the topic REDOX. Each of the case studies, on pages 302 to 311, has been allocated a single page. At the top of each page is the associative map of an individual student as it relates to the topic selected. To facilitate rapid interpretation, areas of the map in which concepts are most readily accessible from one another (threshold levels 1 and 2) have been highlighted in colour, and coloured lines have been added to enclose the strong second-order links represented by level 3. Except for those appearing at the first level, the threshold at which each linkage was added to the map is indicated by the numbers 2, 3, 4 or 5 on the line representing the link. The corresponding values of the relatedness coefficients are summarised below:

Threshold level 1 -	relatedness	coefficient	>	0.50
" " 2 -	"	"	>	0.40
" " 3 -	"	"	>	0.30
" " 4 -	"	"	>	0.20
" " 5 -	"	"	>	0.10

Immediately under each map are comments relating to features of general interest, and these are followed by brief notes relating the student's attempts at each of five items to particular features of his associative map. In these notes, reference is made to data from the self-reporting forms including some indication of the confidence which the student reported concerning the correctness of his answer.

overlapping with the topic REDOX. Each of the case studies, on pages 302 to 311, has been allocated a single page. At the top of each page is the associative map of an individual student as it relates to the topic selected. To facilitate rapid interpretation, areas of the map in which concepts are most readily accessible from one another (threshold levels 1 and 2) have been highlighted in colour, and coloured lines have been added to enclose the strong second-order links represented by level 3. Except for those appearing at the first level, the threshold at which each linkage was added to the map is indicated by the numbers 2, 3, 4 or 5 on the line representing the link. The corresponding values of the relatedness coefficients are summarised below:

Threshold	level	1 -	relatedness	coefficient	>	0.50
"	"	2 -	"	"	>	0.40
"	"	3 -	"	"	>	0.30
"	"	4 -	"	"	>	0.20
"	"	5 -	"	"	>	0.10

Immediately under each map are comments relating to features of general interest, and these are followed by brief notes relating the student's attempts at each of five items to particular features of his associative map. In these notes, reference is made to data from the self-reporting forms including some indication of the confidence which the student reported concerning the correctness of his answer.

1. A dilute solution of sulphuric acid is electrolysed between platinum electrodes.

(a)(i) or (101) State the name of the product formed at the anode.

(ii) or (102) State the name of the product formed at the cathode.

(d) or (104) Explain how the method of conduction of an electric current through dilute sulphuric acid differs from that through a copper wire.

4(b) or (403) At which electrode, if either, are ions reduced during electrolysis?

(c) or (404) A simple cell is formed by dipping strips of zinc and copper into an aqueous solution of an electrolyte. When a small current is drawn from the cell which strip of metal will be the positive terminal?

---

2. The table below gives the number of electrons present in six neutral atoms labelled E to X.

	E	G	J	L	M	X
Number of electrons	6	9	14	16	16	19

Use the letter at the top of a column to represent the element or one atom of the element.

(a) or (201) Which two different elements are in the same group of the periodic table?

(b) or (202) Which atoms, if any, are isotopes.[1]

(c) or (203) Which element is the best reducing agent?

(d) or (204) Which element is a halogen?

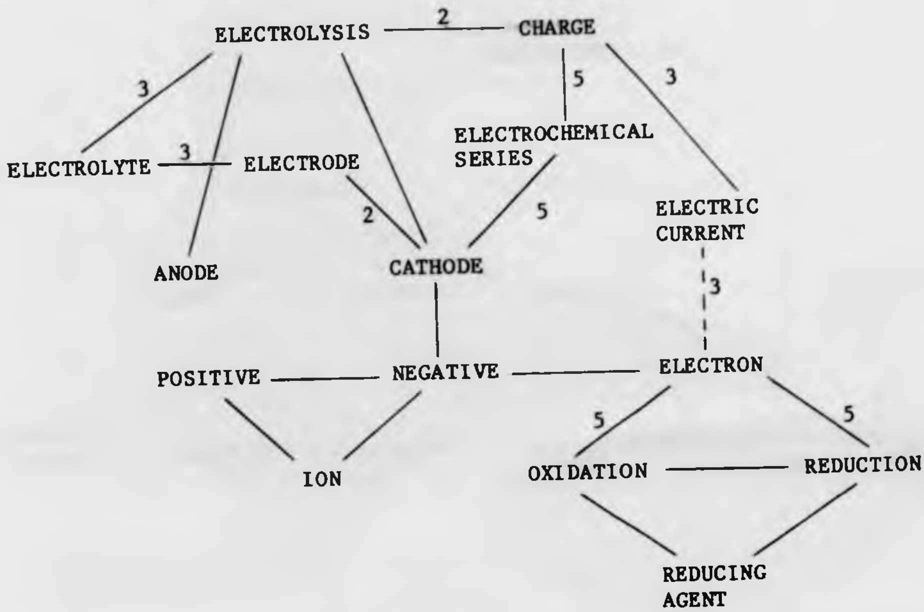
(e) or (205) Derive a formula for the compound formed when X reacts with M.

Figure 7.7 Test items on the topics ELETROLYSIS and ATOMIC STRUCTURE/ PERIODICITY

---

1. In the "standard" version of item 2(b) or (202), the table was expanded to include data relating to the mass numbers of each atom.

Case study 1: ELECTROLYSIS; student 060 (perfect score)



In many individual maps CATHODE appeared more "meaningful" than ANODE (see also Appendix C). However, the latter might possibly be defined in PS situations by contrast with the former. The unusual linkages of ELECTROCHEMICAL SERIES (ECS) and of CHARGE to ELECTROLYSIS, may also be noted.

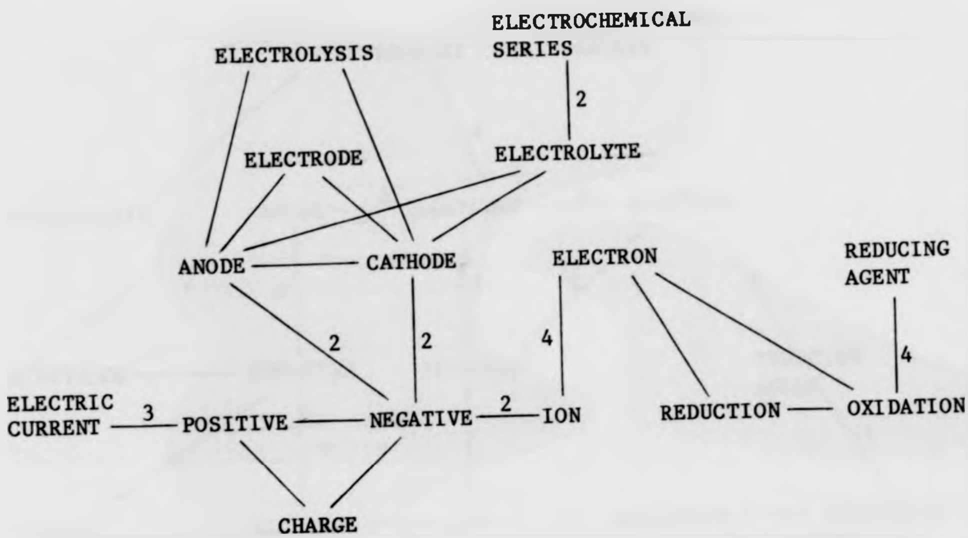
Items 101/2. Confidently recalled as electrolysis of water (strategy S1); assigned gases to electrodes using "opposite charges attract" rule and correct signs on ions and electrodes; reference to "anode positive" presumably by contrast with CATHODE-NEGATIVE in map.

Item 105. Confidently compiled recalled information (S1); current as flow of electrons in metal wire and attraction of ions to, and exchange of electrons at, electrodes; consistent with map reflecting relevant defining/describing relationships.

Item 403. Initially confused over direction of electron transfer in reduction (note weak, symmetrical links from redox cluster to ELECTRON); but confident solution worked out (S2) from defining/describing propositions consistent with map.

Item 404. Despite "augmented (relevant)" version giving data regarding the electrochemical series, unable to solve and forced to guess (SO'); unusual (and weak) linkage of ECS appears to reflect absence of propositional knowledge usable in this problem.

Case study 2: ELECTROLYSIS; student 004 (perfect score)



Note the linkage of both ANODE and CATHODE to NEGATIVE and the weakish linkage of the latter to ELECTRON through ION; also the dubious linkages of ELECTRIC CURRENT and ELECTROCHEMICAL SERIES (ECS).

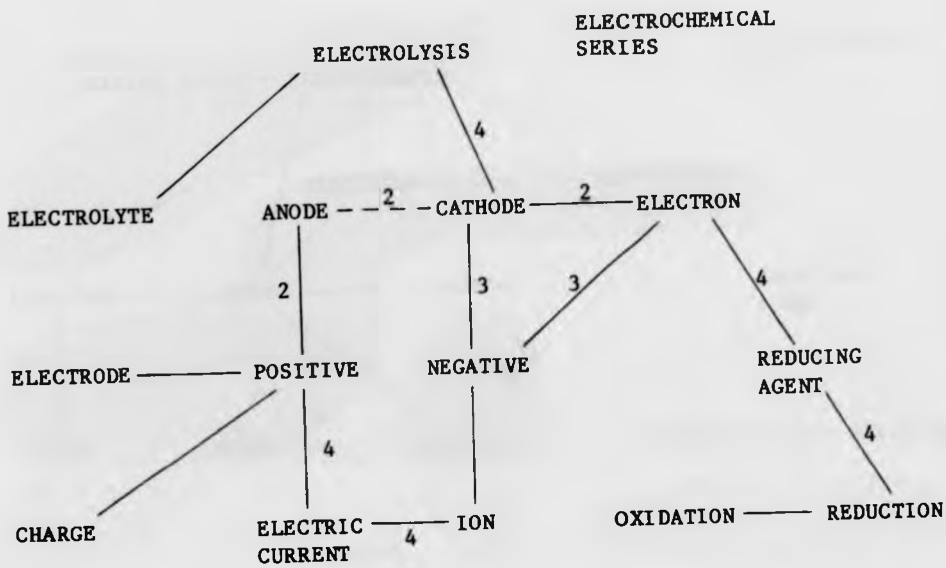
Items 101/2. Solved by specific recall of the particular example (S1); map irrelevant.

Item 105. Stated "remembered H^+ ions in solution and connected to free moving electrons in metal" (S1); note first reference to IONS consistent with closer linking of ION and weaker and more distant linkage of ELECTRON and ELECTRIC CURRENT in map.

Items 403/4. Direct recall (S1)?; map irrelevant but student appeared to confuse voltaic cell effect with electrolysis. In this and other cases the absence of particular concepts relating to the former limited the interpretive power of the map.

It was noted that a perfect score on a limited number of items may be coinsistent with an associative map which contains a number of apparent weaknesses.

Case study 3: ELECTROLYSIS; student 027 (score 1 out of 5)



Note the limited structure, particularly at higher levels of linkage; also the isloation of ELECTROCHEMICAL SERIES (ECS).

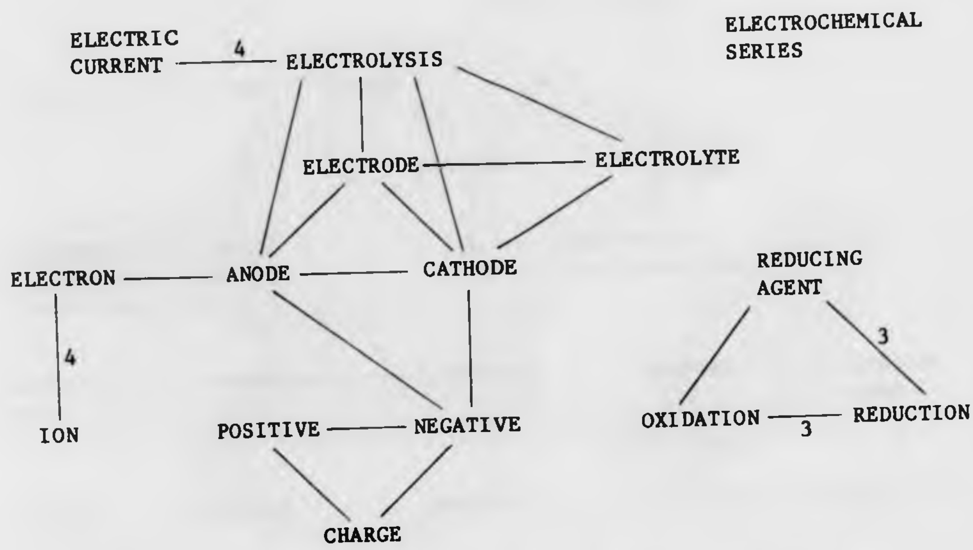
Items 101/2. Errors due to uncertain misrecall (guess ?) of sign of H ion (error E1; map suggests student may tend to associate ION with NEGATIVE and ELECTRODE with POSITIVE) and to seeking (guessing) a product from the sulphate ion (selection of wrong approach, E2'; latter consistent with isolation of ECS).

Item 105. Low confidence guess relating only to the "freer" movement of "charged molecules" in sulphuric acid; there are no strong links/ chains of links in the map between ELECTROLYSIS and ION or between ELECTRIC CURRENT and ELECTRON.

Item 403. Fails to recall (EO) and cannot work out so guesses (SO'); (chain of linkages REDUCTION - REDUCING AGENT - ELECTRON - CATHODE might conceivably account for correctness of "guess").

Item 404. Makes low-confidence, confused attempt to grope towards an answer (serial processing, Sp); attempt is consistent with isolation of ECS.

Case study 4: ELECTROLYSIS: student 138 (zero score)



Note the very high proportion of high level linkages, some of which may be poorly discriminated (ANODE to NEGATIVE and to ELECTRON); also the relative isolation of ELECTRIC CURRENT and ION and the complete isolation of the redox cluster and ELECTROCHEMICAL SERIES (ECS).

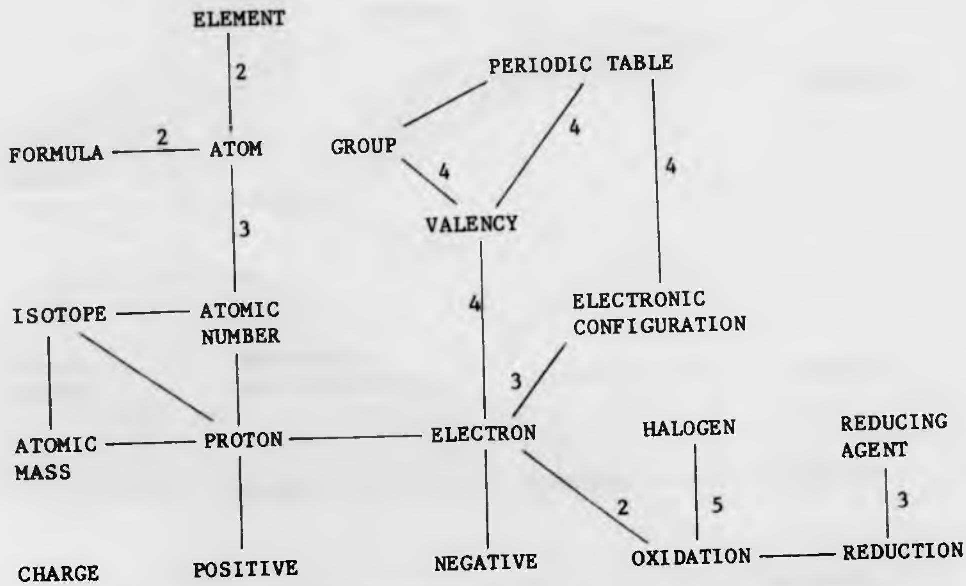
Items 101/2. Low confidence guesses (SO'); cannot recall anything about "electrolysis of sulphuric acid" so map irrelevant.

Item 105. Omits (SO); unable to attempt because "haven't covered electrolysis in solid substances"! Consistent with relative isolation of ELECTRIC CURRENT and the apparently poorly discriminated links to ELECTRON and ION (which do not relate to normal defining or describing relationships).

Item 403. Guesses (SO'); consistent with isolation of redox cluster.

Item 404. Guesses (wrongly) with correct reference to the relative reactivity of zinc and copper (SO'); consistent with isolation of ECS but map unhelpful due to absence of particular concepts such as CELL, ELECTRODE POTENTIAL, COPPER, ZINC etc.

Case study 5: ATOMIC STRUCTURE/PERIODICITY; student 006 (perfect score)



Note the strong cluster of concepts around PROTON and the central positions of VALENCY and ELECTRONIC CONFIGURATION (E CONFIG) linking the periodicity cluster to the atomic structure cluster; also the isolation of the small cluster related to ATOM and the unusual, though not inexplicable, position of HALOGEN.

Item 201. Moderately confident solution working from number of ELECTRONS to E CONFIG and VALENCY (S2); consistent with map.

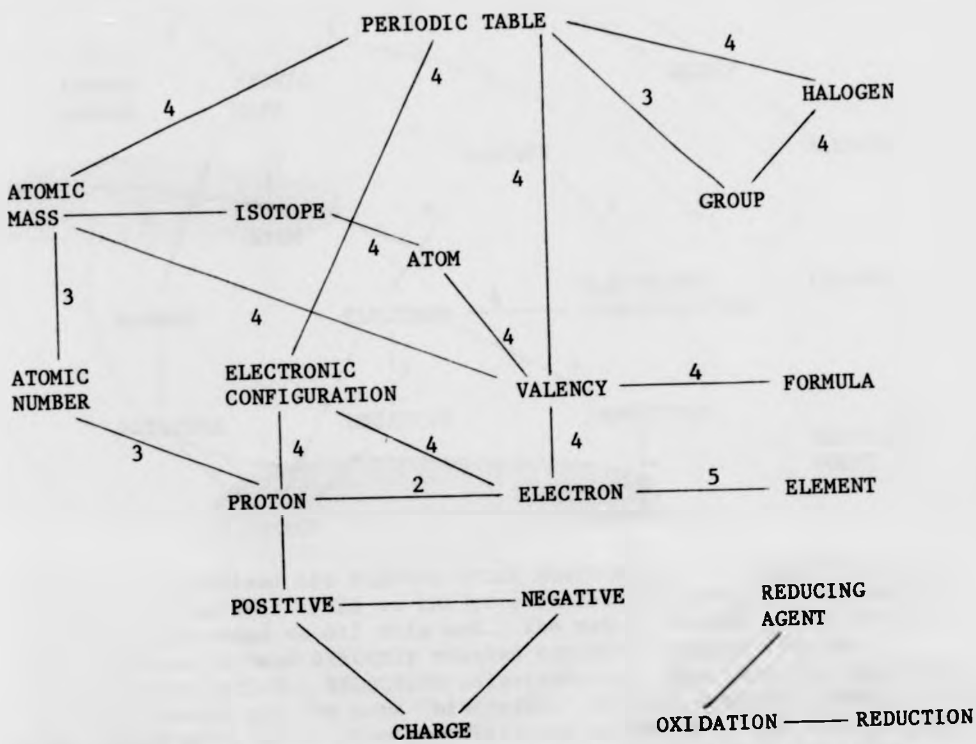
Item 202. Confident solution from ISOTOPE, via "same number of PROTONS" to same number of ELECTRONS (S2); greater confidence may be associated with stronger linkages involved.

Item 203. Certain of solution based on definition of REDUCTION and working out E CONFIGs (S2; initial processing error E-2 self-corrected); consistent with map.

Item 204. Blind guess, "can't remember what HALOGEN is"; note HALOGEN not linked to PERIODIC TABLE cluster. Element "G" (the halogen), obtained following E-2 "inversion" error during attempt on previous item, may have been recognised as an oxidising agent when this was corrected; subsequent "guess" in present item might, speculatively, have employed HALOGEN - OXIDATION link at an "intuitive" level.

Item 205. Confident solution based on VALENCY obtained from number of ELECTRONS and E CONFIGs (S2). Note that "primitive" FORMULA - ATOM link in map obscures direct FORMULA - VALENCY link at level 4 and that FORMULA is in any case strongly linked to ELECTRON through the map as shown.

Case study 6: ATOMIC STRUCTURE/PERIODICITY; student 005 (perfect score)



A generally well constructed map; note the large number of links at level 4 particularly those joining the PERIODIC TABLE cluster to the ATOMIC STRUCTURE cluster through VALENCY and ELECTRONIC CONFIGURATION (E CONFIG); also the isolation of the redox group.

Item 201. Confident solution using E CONFIGs (S2); consistent with map.

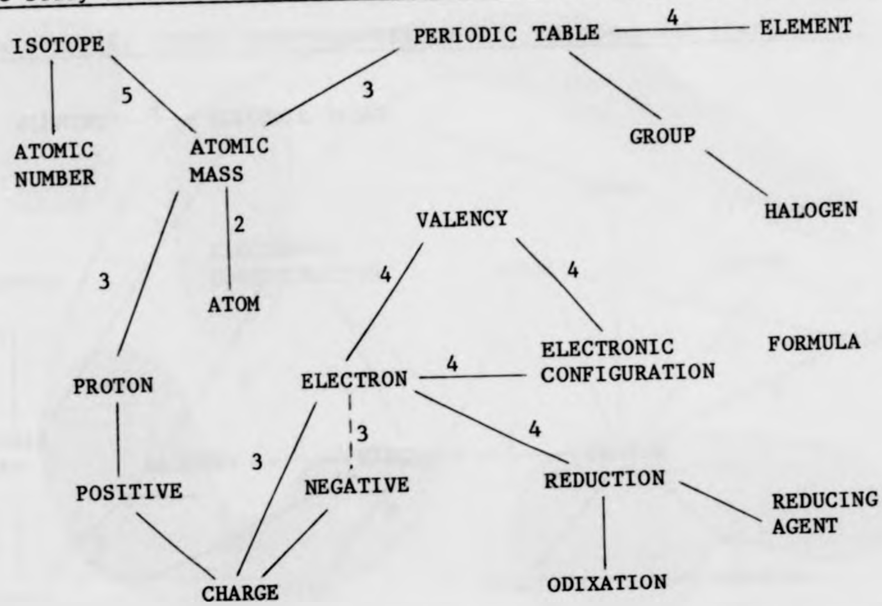
Item 202. Confident solution referring to "same numbers of ELECTRONS and PROTONS" (S2); PROTON logically accessible from ISOTOPE through strong links; solution considered consistent with map.

Item 203. Correct solution including reference to gain of ELECTRONS as REDUCTION (S2), but considered answer "probably wrong"; solution not consistent with isolation of redox cluster, but the latter may be associated with the student's low confidence in her answer.

Item 204. Correct solution consistent with map (S2; only moderate level of confidence appears to reflect uncertainty whether HALOGENS are in GROUP seven).

Item 205. Certain of correct solution based on "filling outer shells" of ELECTRONS (S2); consistent with map.

Case study 7: ATOMIC STRUCTURE/PERIODICITY; student 175 (perfect score)



Student 175 obtained the highest total score on the PS test as a whole as well as a perfect score on the present set of items; she was certain about her solutions to all this set. The map is rather linear but all the links are between directly related concepts (except for the reciprocal OXIDATION - REDUCTION relationship). The relative isolation of ATOMIC NUMBER and the more "historical" link of PERIODIC TABLE to ATOMIC MASS were noted; also the isolation of FORMULA. In practice this student solved most of the items simply by reference to a recalled tabulation of the first 20 elements in the periodic table, identifying the elements in the question by equating the number of electrons with order of appearance in the table, and recalling relevant properties. The systematic nature of the map may have contributed to the probable "reconstructive" aspect of the recall involved but had limited obvious relevance to most particular solutions.

Item 201. Certain answer based on reconstructive recall as described above (strategy S1); map of limited relevance.

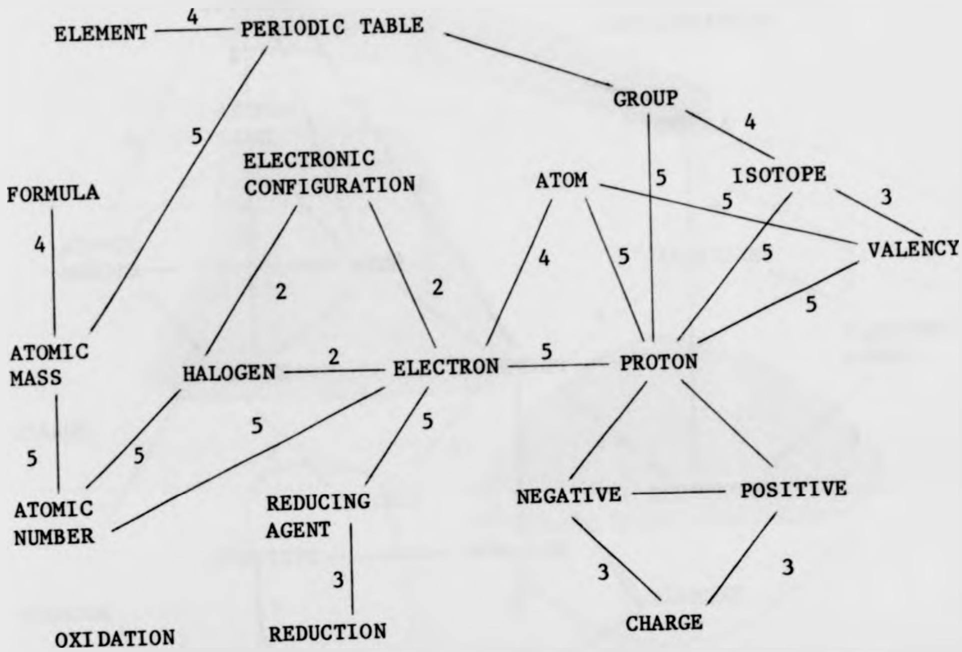
Item 202. ISOTOPES identified as having same number of protons, and therefore electrons (S2); consistent with the assumed logic of the map (although the addition of NUCLEUS and NEUTRON would probably improve detailed interpretability).

Item 203. "REDUCING AGENTs - wants to be oxidised ∴ metal", selects the only metal present in the list using recalled table; strategy difficult to classify satisfactorily but consistent with map.

Item 204. Knows HALOGEN is group 7 and selects from recalled table (S1); map of limited relevance but note strong identification of HALOGEN.

Item 205. Identifies elements from recalled table, recalls valencies and writes formula; strategy difficult to classify satisfactorily; map of limited relevance (note isolation of FORMULA).

Case study 8: ATOMIC STRUCTURE/PERIODICITY, student 002 (zero score)



This map showed little structure after the third, and even the fourth, linkage thresholds; many of the links did not reflect defining or describing relationships (eg PROTON to NEGATIVE and VALENCY, ISOTOPE to GROUP); note also the isolation of OXIDATION. Given the numbers of both ill-discriminated and weak linkages it was not surprising that this student omitted all but one of the items concerned.

Item 201. Omits (SO); "can't see any similarities"; note absence in map of any strong, well discriminated chain of links between GROUP and ELECTRON/ELECTRONIC CONFIGURATION.

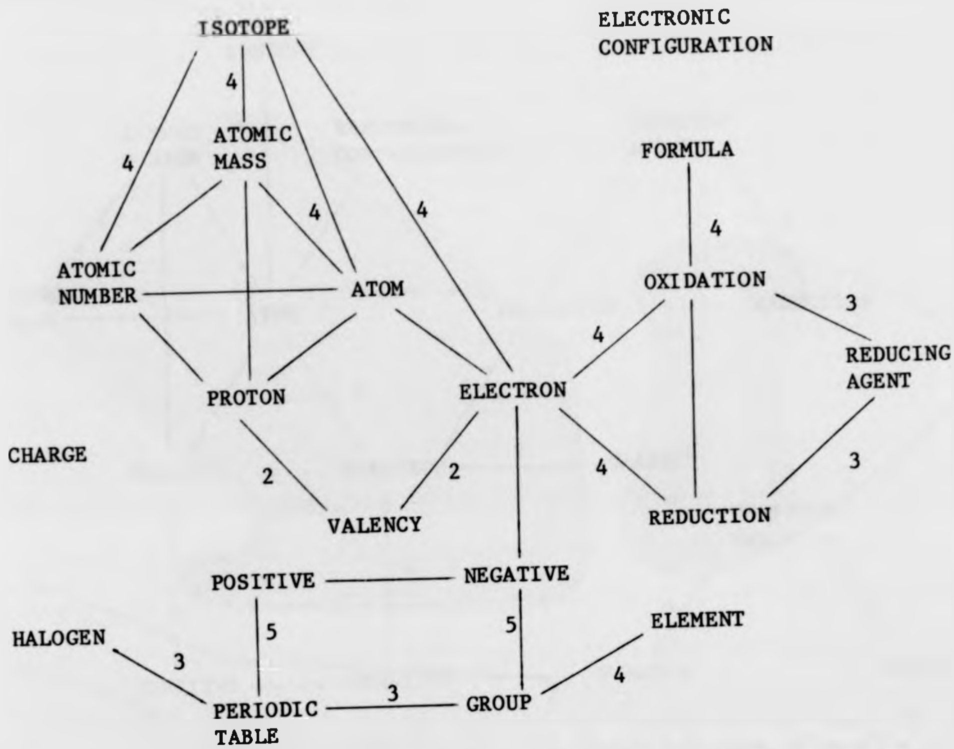
Item 202. Omits (SO); "can't remember what ISOTOPE means"; note absence of associations of ISOTOPE with ATOMIC MASS etc.

Item 203. Omits (SO); "no idea"; even if the REDUCING AGENT - ELECTRON link well discriminated, no useful linkage of ELECTRONIC CONFIGURATION to concepts relating to periodicity.

Item 204. Guesses M as halogen (SO'), rationalising with doubly incorrect reference to "full outer shell". Note strong, but obviously unhelpful, association of HALOGEN to ELECTRONIC CONFIGURATION which may account for this attempt; also absence of links likely to define HALOGEN in relation to periodicity.

Item 205. Omits (SO); "not enough knowledge to answer"; note isolation from one another of FORMULA, VALENCY and ELECTRONIC CONFIGURATION.

Case study 9: ATOMIC STRUCTURE/PERIODICITY, student 020 (zero score)



Note the quite well structured and strongly linked atomic structure cluster and the adequate redox cluster; also the idiosyncratic linkage of the PERIODIC TABLE cluster to POSITIVE and NEGATIVE, of FORMULA to OXIDATION and of VALENCY to PROTON, and the isolation of CHARGE and, particularly, ELECTRONIC CONFIGURATION (E CONFIG). This student attempted the augmented (irrelevant) version of these items in which additional data regarding mass numbers was provided, and made no reference at all, in either answers or accompanying notes, to numbers and arrangements of electrons.

Item 201. Omits (SO) after failing to recall how mass numbers relate to the periodic table (EO); absence of any attempt to use data on number of electrons is consistent with map (see general note above).

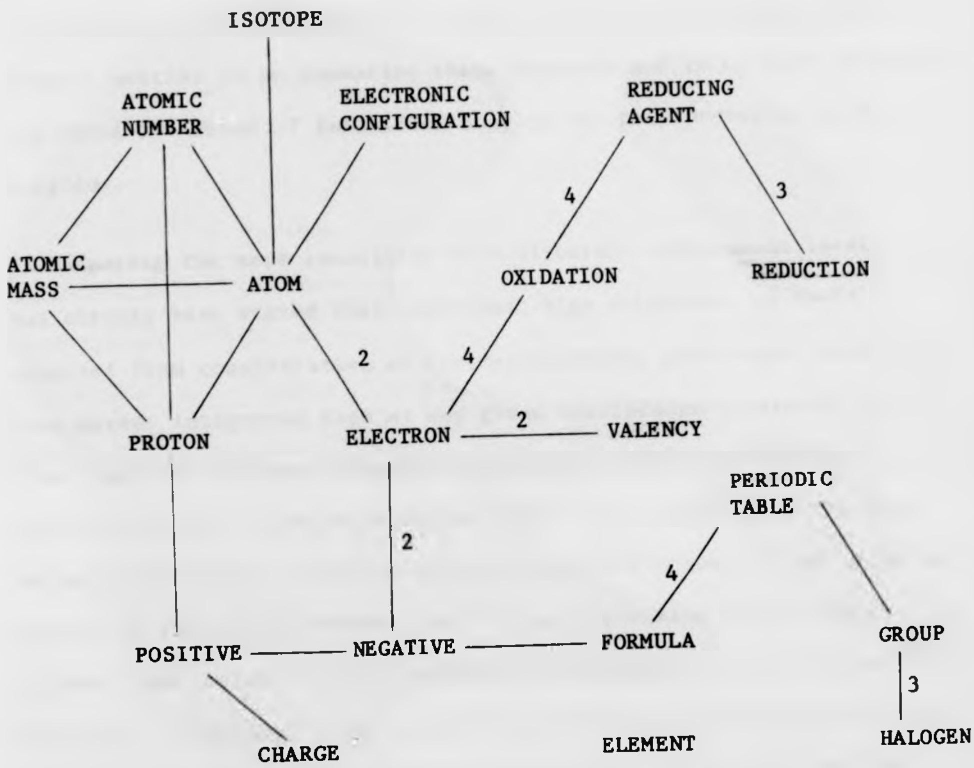
Item 202. Low confidence guess based on perceived regularity in mass number data alone (SO); numbers of electrons ignored (see above).

Item 203. Omits (SO) after failing to connect reducing agents and mass numbers; numbers of electrons ignored (see above).

Item 204. Omits (SO); "I don't understand how it is worked out."

Item 205. Low confidence guess based on ratio of mass numbers (SO'); numbers of electons ignored (see above).

Case study 10: ATOMIC STRUCTURE/PERIODICITY; student 044 (zero score)



Though strongly linked and well structured in part, note the poorly discriminated strong associations of ISOTOPE and ELECTRONIC CONFIGURATION (E CONFIG) to ATOM and of NEGATIVE to FORMULA; also the relatively weak linkages of the redox cluster, the idiosyncratic nature of the link between the atomic structure and periodicity clusters and the isolation of ELEMENT.

Item 201. Guesses on basis that "numbers are the same" (SO'); also comments "not sure about E CONFIG rules" which is consistent with the poorly discriminated linkage of E CONFIG in the map.

Item 202. Guesses without comment (SO'); note poorly discriminated linkage of ISOTOPE.

Item 203. Guesses (SO'); inabilty to use data concerning numbers of electrons consistent with earlier comments regarding map.

Item 204. Guesses (SO'); cannot use knowledge of HALOGEN in GROUP 7 in absence of useful propositions regarding E CONFIGs.

Item 205. Omits (SO); no way to proceed (see notes above).

7.4.4 An interpretive summary of the case studies. Some generalisations regarding the cognitive maps of individual students were discussed in the course of introducing the case studies. The purpose of the present section is to summarise these findings and to consider evidence regarding the value of individual maps in the interpretation of PS episodes.

In comparing the maps associated with different achievement levels it has already been stated that individual high achievers, as would be expected from consideration of the corresponding group maps, tended to have better integrated maps at any given relatedness threshold, with fewer isolated concepts, than low achievers. Although individual differences were often quite marked, the maps of high achievers also tended to show less variation than those of low achievers and to be more similar to the group average map. The most striking differences, however, were related to the **quality** of the associations reflected in the maps. In general those of the high achievers tended to link closely related concepts and were readily interpretable in terms of logical chemical relationships, while the maps of low achievers tended to contain a number of apparently indiscriminate associations. At the individual level, even the high achievers did not mirror in detail the a priori subject-matter structures suggested by the author on the basis of a consideration of normal teaching practices (Appendix C). However, where linkages differed they were generally readily interpretable. Thus the fairly common linkage of VALENCY to ELECTRON, and of PERIODIC TABLE to ATOMIC MASS, seemed likely to reflect useful learnt relationships. Indeed, it has already been pointed out that the a priori structure presented makes no claim to be a unique representation of the chemical relationships involved (although it proved to be quite a good predictor of the average map). By contrast, idiosyncratic links in the maps of low achievers such as those joining PROTON to NEGATIVE and VALENCY, PERIODIC TABLE to POSITIVE, or FORMULA to OXIDATION, appeared to have limited meaning. And whilst correct propositions could be suggested to account for such linkages as those between ELECTRON and, in different maps, ANODE, CATHODE and ION, these were also judged most unlikely to reflect useful learnt relationships. Similarly, the virtually complete inter-linkage of the concepts ELECTROLYSIS, ELECTROLYTE, ELECTRODE, ANODE and CATHODE in case study 4, was assumed to represent merely vague associations of limited utility (associations of this sort have been linked with early stages of learning (Preece, 1977)).

Before going on to consider the explanatory value of individual maps, two points must be made in relation to the issue of poorly discriminated associations (that is to say, associations which appeared to be merely idiosyncrataic or vague and unlikely to represent significant, correct relationships). Firstly, although such associations were rare in the maps of high achievers, they were not entirely absent. In general, however, they tended to affect parts of maps which were not involved in the particular problem solutions adopted by the student concerned and thus did not influence his PS performance. Secondly, the identification of well and poorly discriminated associations depended to some extent on subjective judgements (whether or not reference was made to an a priori structure). Although these could be rationalised with some confidence, the use of a battery of acceptable relationships, possibly in conjunction with an examination of the raw word association data as well as the final map, might provide a basis for the development of more objective criteria in future studies.

The potential value of individual associative maps in the interpretation of particular PS episodes was well demonstrated by the case studies. PS attempts always appeared to be consistent with the maps insofar as the latter were relevant to the particular solution route selected; and where routes were chosen which bypassed the map, it was often possible to interpret such a choice of route by reference to defects in the map. In a number of cases the failure of standard S2 (propositional reasoning) solutions was associated with the isolation of an essential concept or the absence of an essential link. The isolation of ELECTRONIC CONFIGURATION in case study 9 was a case in point, as was the lack of any meaningful chain of associations linking the same concept to the periodic table cluster in case studies 8 and 10. Conversely, the linkage of ELECTRONIC CONFIGURATION and VALENCY as bridges between the periodic table and atomic structure clusters in case study 5 was associated with correct and confident S2 solutions to relevant items. In only one case was a correct solution obtained which appeared to make use of propositional knowledge which was not reflected by a link in the map; it was of interest that the student concerned considered that his answer was "probably wrong" (case study 6, item 203). It was, in fact, notable that a student's confidence in his answer (as recorded on his self-reporting form) was a good reflection of the apparent adequacy or otherwise of the relevant parts of his map. Finally, the maps appeared to offer interpretations of a number of cases of omitting and guessing. Not only were these invariably associated with relevant defects in the student's map, but it was sometimes possible to follow a chain of associations which appeared to account, if rather tentatively, for a particular guess (eg case study 3, items 101 and 403).

A particular limitation on the interpretive power of the individual associative maps was the relatively high incidence of solution routes which bypassed the maps. Thus, in cases where students relied on the recall of specific examples, or employed propositional reasoning involving specific concepts which were not included in the WAT, the maps tended to be partly or wholly irrelevant. Case study 2 illustrates how, in such cases a perfect score on a limited set of items may arise despite poorly structured and idiosyncratic maps. For this reason an attempt to use individual maps to "predict" whether or not particular students would be found, on subsequent inspection of their answer scripts, to have succeeded in solving particular items, met with rather limited success and added nothing to the picture already presented in detail in the case studies. Despite this limitation, however, the success of the individual maps in providing a rational basis for interpreting PS episodes within their purview, together with the systematic differences observed between the maps of high and low achievers, clearly lent support to the hypothesis that the organisation of semantic knowledge in LTM is an important factor in determining PS performance.

7.5 A discussion of the findings

The analysis of students' PS protocols during the first phase of the study suggested that there were few differences between high and low achievers as regards the sorts of strategies which they adopted and the sorts of errors which they made. The results of the cognitive mapping exercise described above were consistent with the hypothesis, made at that time, that differences in PS behaviour were attributable to differences in the organisation of knowledge in LTM. In reaching such a conclusion, the associative mapping technique used to elucidate students' cognitive structures obviously played a crucial role. In the short discussion which follows, it will be convenient to consider together both the findings and the techniques which helped to elicit them.

Both the average and the individual associative maps
organisation of the problem-relevant knowledge of hi
differed particularly in the number and strength of
between directly related concepts, that is to say, b
standing in important defining and describing relati
another. It seems natural to assume that, in genera
represented correctly the relationships referred to.
self-consistent picture in which high mean associati
directly related concepts are associated with mean g
solving relevant problems. External validation for
provided by the convergence of associative and subj
structures. However, the individual maps of low ac
characterised by the high incidence of idiosyncrati
concepts which are not directly related. While the
relationships represented by such linkages remained
clear that they reflected, at best, inadequate unde
concepts concerned and of their interrelationships.
maps (that is to say maps showing a general resemb
matter structure) tended to be associated with str
confident solutions, idiosyncratic maps were often
confused attempts and low self-confidence.

A high proportion of the errors identified during the first phase of the study were associated with difficulties in recalling the knowledge required. Such knowledge might be regarded as consisting of a succession or chain of propositional links between concepts, starting with the information given in the problem and leading directly, or through logical manipulation, to an information state recognised as representing the answer. The average map of the low achievement group, Both the average and the individual associative maps suggested that the organisation of the problem-relevant knowledge of high and low achievers differed particularly in the number and strength of associative linkages between directly related concepts, that is to say, between concepts standing in important defining and describing relationships to one another. It seems natural to assume that, in general, such linkages represented correctly the relationships referred to. This leads to a self-consistent picture in which high mean associative linkages between directly related concepts are associated with mean group success in solving relevant problems. External validation for this picture is provided by the convergence of associative and subject-matter structures. However, the individual maps of low achievers were characterised by the high incidence of idiosyncratic linkages between concepts which are not directly related. While the exact nature of the relationships represented by such linkages remained unknown, it seemed clear that they reflected, at best, inadequate understanding of the concepts concerned and of their interrelationships. Thus while "normal" maps (that is to say maps showing a general resemblance to the subject-matter structure) tended to be associated with straightforward, confident solutions, idiosyncratic maps were often associated with confused attempts and low self-confidence.

A high proportion of the errors identified during the first phase of the study were associated with difficulties in recalling the knowledge required. Such knowledge might be regarded as consisting of a succession or chain of propositional links between concepts, starting with the information given in the problem and leading directly, or through logical manipulation, to an information state recognised as representing the answer. The average map of the low achievement group, as compared to that of the high achievers, showed where such knowledge chains might be likely to fail through the absence, or the limited relative immediacy, of critical links. The associative maps of particular individuals often demonstrated this still more clearly, and also indicated where idiosyncratic and potentially misleading associations might intervene. Unfortunately, the limited information available from the students' self-reporting forms, as compared to that obtainable through the collection and analysis of spoken protocols, made it possible to analyse particular PS episodes only at a rather speculative level.

The present study as a whole has tended towards the conclusion that differences in the ability to answer specific examination items correctly are dependent to a major extent, not on strategies or IQs, but on the presence or absence of appropriately organised knowledge in cognitive structure. Global success in examinations thus appears to be a function of the interaction of a personal cognitive map with a particular set of test items. It has been noted, in the discussion above, that the mean associative map for a group covers a wide range of such personal maps, reflecting relatively wide differences in the immediacy of given associations. While a few students may conform reasonably closely to the mean map, there are many individual differences which may or may not impede the solution of particular problems. This is entirely consistent with the topic-specific nature of individual examination performance. It suggests that an individual's performance may be an "average" which hides significant weaknesses in specific areas (or perhaps as regards particular concepts), the "good" chemistry examinee being the one with fewest such areas. This analysis also suggests that poor performance in chemistry may be traced to a series of quite specific knowledge defects. It seems likely that those occurring early in learning may also limit possibilities for the adequate linkage of later ideas. This leads to the speculation that the (early) identification of such specific defects might indicate specific remedial teaching and thus lead to drastically improved performances. The ideas raised in this paragraph, and the possible future role of word association methods in work of this kind, are discussed further in Chapter 8.

CHAPTER 8

A SUMMARY OF THE STUDY AND A DISCUSSION OF SOME OF ITS IMPLICATIONS

The first section of this chapter is devoted to a brief review of the study and its findings. The implications of these, in relation to different aspects of the work, for the teaching and examining of chemistry are then considered and some suggestions are made regarding further research. A short concluding section completes the report.

8.1 A review of the study and its findings

After five years of secondary schooling, a student's achievements in a particular subject are generally reported in the form of a single examination grade. However, whilst the objectives of such examinations are often stated in cognitive terms there appears to have been little empirical research into the cognitive events which actually underlie examination success. In this thesis a study has been described which attempted to investigate, in a broad and exploratory way, the cognitive behaviour of GCE O-level chemistry students in an examination situation. During the first phase of the work a protocol approach was used to generate detailed descriptions of the PS strategies and errors of examination candidates in terms of a simple information processing model. It was found that the students concerned relied heavily on simple, memorised knowledge and that a great majority of the short-answer, non-mathematical examination items employed in the study were answered either by direct recall or by the routine manipulation of given and recalled knowledge. Although the work in this part of the study was not quantitivate, it was estimated that over 80% of the errors recorded could be directly attributed to recall difficulties, and only a minority to "processing" errors. The PS strategies and errors of the more successful candidates differed little from those of the less successful, and the performance of the students involved in the study was unrelated to general intelligence as measured by the AH4 test. As a result of this phase of the work, it was hypothesised that differences in PS performance might be attributable to differences in the organisation of relevant knowledge in a student's LTM.

The second phase of the study fell into two parts. The first of these was concerned with exploring one aspect of the relationship between performance in examinations and performance in more realistic chemical situations. An analysis-of-variance design was used to investigate the influence of the level of information provided in an examination item on the PS performance of students. It was found that, in general, performance tended to fall as the level of information was raised, irrespective of the relevance of the information to the task in hand.[1] On the basis of the detailed analyses, and of supplementary data from self-reporting forms, it was concluded that students were generally uncritical in assessing the demands of PS tasks and that they tended to process available data indiscriminately. This must be considered to have serious implications for PS performance in everyday situations in which levels of information may not be as closely controlled as they are in examination items.

The final aspect of the study was concerned with investigating the hypothesis, generated as a result of the first phase of the work, that differences in the performance of examination candidates could be largely attributed to differences in the organisation of problem-relevant knowledge in a student's LTM. For the purpose of this

1. There was also a weak counter-trend associated with increasing levels of relevant cuing.

investigation, word association methods were used to map students' semantic networks. The hypothesis was broadly confirmed; the cognitive maps of high achievers tended to be more extensive and better integrated than those of low achievers and to contain better discriminated linkages (that is to say, linkages between concepts which are closely related in the logical structure of chemistry). Moreover, individual cognitive maps were found to have some explanatory potential in relation to particular PS episodes.

In the sections which follow, the findings summarised above are discussed in relation to their implications for the practice of science teaching and for the conduct and interpretation of examinations. Priorities for further research are also considered.

8.2 The educational implications of the protocol analysis and word association studies

The protocol analysis and cognitive mapping aspects of the study strongly suggest that modern O-level chemistry examinations discriminate between students, not on the basis of their cognitive skills as such, but according to the way in which their knowledge of chemistry is internalised as a semantic network in LTM. On the surface, this may seem to suggest a mis-match between what examinations set out to test (that is to say, not only knowledge but cognitive abilities such as comprehension, application and evaluation) and what they actually assess (namely the quality of students' semantic structures). However, it is suggested that the sound organisation of semantic knowledge in LTM is in fact the principal prerequisite for the exercise of such cognitive abilities.

The mediating role of LTM between learning and behaviour has been highlighted in a number of studies relating to mathematics and science education (Mayer and Greeno, 1972, 1974; Mayer, 1975; Gagné and White, 1978; Atkin, 1978) and the importance of well organised, conceptual knowledge in solving problems in science has also been referred to in literature cited earlier in the thesis (eg Wason and Johnson-Laird, 1972; Bhaskar and Simon, 1977; Larkin, 1980). Moreover, structural knowledge of relevant content is seen by some educational psychologists to be of more importance in learning and performance than stages of intellectual maturation and the development of associated levels of cognitive functioning (eg Novak, 1976, 1978; Driver, 1978). Similarly, factor analysis studies of examination performance have focused attention on the importance of topic knowledge rather than cognitive skills (eg Kempa and L'Odiaga, 1983). The principal finding of the present study, as reported above, is fully consistent with these ideas.

The central role of LTM in problem-solving, as it was reflected in this study, can be quickly reviewed. In terms of the information processing model adopted, the student's perception of the problem and his consequent selection of an approach to solving it involved immediate reference to LTM. Only insofar as the relevant structures stored there were both accessible to the cuing available and adequate for the comprehension of the PS task, was progress possible. Complete misunderstanding of the question accounted for only a small number of failures to solve items. However, **inadequate** understanding was at the root of many subsequent errors. Most attempted solutions were based either on direct recall of the solution or on the routine, logical manipulation of given and recalled information (propositional reasoning). Other attempts represented a limited range of responses to the failure of these two basic strategies and included the use of recalled analogies, groping around to recall related information ("serial processing") and "guessing". The generation of worthwhile hypotheses and their systematic evaluation was very rare. Most PS breakdown was associated directly with inadequacies in the information stored in LTM and in particular with failure to locate essential information. Errors of this sort were associated with cognitive maps which were poorly integrated and which tended to contain poorly discriminated linkages (that is to say, linkages between concepts which are not closely related and which were therefore assumed to have little useful information content). Processing errors were of only limited importance and appeared to be associated with overloading the limited capacity of STM. However, Johnstone (1981) has pointed out that these, and indeed errors associated with inadequate comprehension of the problem, may be reduced if the ideas involved can be so organised in memory as to facilitate their manipulation in meaningful "chunks". This would suggest that even breakdowns in processing and task perception might, in fact, ultimately be attributed to deficiencies in the structuring of information in LTM.

8.2.1 **Implications for teaching.** If examination performance, and in particular failure in examination tasks, are closely related to relevant aspects of students' semantic networks, then teachers will obviously want to take account of this in planning their students' learning. More particularly so, since most of the studies cited earlier also suggest that future learning and success in "near" and "far" transfer tasks depend on the same structures.[1] Evidence from such studies leads to the conclusion that teaching should be directed towards making the structure of the subject matter explicit, emphasising major concepts and their inter-relationships, stressing unifying principles and establishing firm

1. Success in "near" transfer tasks has been associated with the level of the "internal connectedness" in a body of knowledge, and "far" transfer with the level of the "external connectedness", or integration, of such knowledge with other aspects of cognitive structure (Mayer, 1975).

links with immediately relevant previous learning and with wider aspects of past experience. While many teachers would probably accept the value of such attempts to make learning as "meaningful" as possible it may not, in practice, be easy to integrate such methods with pupil-centred, discovery approaches to the teaching of science. However, the importance of ensuring that students do not build inadequate conceptual frameworks is further highlighted by recent reports regarding the stability of such "naive conceptions" even amongst supposedly sophisticated students (Nussbaum and Novick, 1981; Vincent, 1981) and Pickthorne (1983) has drawn attention to the way in which such "error factors"[1] (which he believes are often based on over- or under-generalisations of a concept or rule) may act to block learning and lead to whole classes of further errors. Such findings direct attention to the likely value of the early detection and correction of inadequate conceptual frameworks and to the need to develop a "Science of Science Teaching" which will take into account the importance of science content, and of optimising the quality of students' acquired semantic networks.

8.2.2 **Implications for examining.** It has recently been reported that the GCE O-level science examinations of a number of boards have shown few changes in question style over the past 20 years (except for the introduction of multiple-choice sections) and more particularly that "very few questions are designed to test the pupils' ability to apply knowledge and understanding of scientific principles" (Wilby, 1983). This is despite the claim that, in addition to "knowledge", such examinations set out to assess, for example:

---

1. The term "error factor" was coined by Harlow (1949) to refer to systematic but incorrect ways of responding to problem situations. Pickthorne's development of this concept enables error factors to be identified with defects in a semantic network.

"Comprehension, application and evaluation:

(a) the ability to understand and interpret scientific information presented in verbal, numerical or graphical form and to translate such information from one form to another.

(b) the ability to explain familiar phenomena in terms of relevant models, laws and principles.

(c) the ability to select and apply known laws and principles to given situations.

(d) the ability to think clearly about given data.

(e) the ability to recognise mistakes and misconceptions."

(Joint Matriculation Board 1979, p 279)

These abilities were reflected in the protocol study by items which students attempted to solve largely by recall and by the logical manipulation of given and recalled information. The incidence of "productive" PS, involving the application of scientific knowledge in novel ways, was very low. Of the attempts recorded, only 5% (labelled S3, hypothesis and evaluation) appeared to involve rational thinking of a kind going beyond familiar routines. Furthermore, only 30% of such attempts resulted in correct solutions. As was observed in Chapter 4, students could hardly be blamed when they reported that they would tend, in an examination, to omit items which appeared to demand such treatment. Nevertheless, insofar as productive PS may be felt to represent a valuable skill which ought to be within the compass of many students, it would seem desirable to increase, in examinations, the proportion of items making such demands. Indeed, only if such items can be seen to represent a worthwhile potential source of marks, would teachers be likely to help the generality of their students to develop the confidence to attempt them.

A second implication relates more fundamentally to the nature of that which is assessed by current science examinations. While the value of testing abilities such as those listed above is not questioned, it may be reiterated that a representative set of items actually discriminated between students principally in terms of the characteristics of their semantic networks. This appears to suggest, prima facie, that these abilities depend heavily upon the quality of such networks in relation to the topic areas covered. It also raises the question of whether the outcome of a student's learning might not validly and conveniently be assessed simply with reference to such structures through the further development of cognitive mapping techniques. Whether or not this would be a desirable alternative to current examination practice is, of course, a different question.

Finally, an important implication of the findings for assessment relates to current interest in graded tests, grade-related criteria and profile reporting. Insofar as the skills involved in answering chemistry examination questions are concerned, the study strongly suggests that these are a function of the organisation of knowledge in a student's semantic network. This, and evidence from other sources, points in turn to the content-related, not to say topic-specific, nature of such skills. It would therefore appear unwise to develop performance criteria without reference to content, or to give the impression, in reporting a student's achievements, that the demonstration of a set of skills in one content area necessarily implies the possession of the same skills in all content areas.

8.3 The educational implications of the task formulation (level of information) study

It was found that students appeared to be undiscriminating in their approach to the information provided in examination items. Thus the provision of information beyond the minimum required to generate an answer, even in some instances where such information was considered necessary to avoid possible ambiguities in the task, tended to lead to confusion and to reduced PS performance. The confusion was associated with a tendency to try to make use of all the information provided irrespective of its redundancy or even relevance. While such tendencies might, to some extent, reflect students' expectations regarding the information provided in examination items, it was clear that in many cases it also reflected an inadequate perception of the task and its demands.

8.3.1 **Implications for teaching.** The role of LTM structure in understanding an item was considered in Chapter 2 (Section 2.3) and has been referred to at a number of points throughout this report. When interpreting the findings of this phase of the study in Chapter 6, it was suggested that the failure of students to discriminate between essential and non-essential information ultimately reflected inadequacies in the way in which relevant knowledge was organised in LTM. As regards the implications of this for classroom practice, it reinforces the case made earlier in this chapter (Section 8.2.1) for teaching based on theories of learning which incorporate appropriate models of cognitive structure. This view finds direct support in experimental work referred to earlier in which the outcomes of teaching based on an Ausbelian model, and designed to lead to "meaningful" learning (that is to say, learning which is not only well structured internally but also well integrated with the existing structural organisation of LTM), were compared with those resulting from more conventional teaching. It was found, inter alia, that students taught with structural outcomes in mind were significantly better at identifying unanswerable questions, that is to say, at identifying the demands of a task (Mayer and Greeno, 1972).

In addition to the implication that teaching should be planned with structural outcomes in mind, there is also perhaps a more obvious implication that students should be **taught** to assess task demands and available information more critically. This might conveniently be done in the context of teaching examination and general PS skills. Unfortunately such courses are rarely found in schools, although they have often been advocated (eg Wooley, 1979).

8.3.2 **Implications for examining.** A recent newspaper article, concerning an as yet unpublished report of research commissioned by the Department of Industry, suggests that modern technological applications are very poorly reflected in GCE (and other) science examinations and that few questions demand the application of scientific principles in realistic situations (Wilby, 1983). Such articles reflect a growing demand that science examinations should be seen to be more relevant to everyday science and technology. The findings of the present study suggest that performance on realistic tasks is likely to fall short of that on existing examination tasks although, insofar as tasks of both sorts appear to depend on the organisation of semantic networks in LTM, the latter are likely to discriminate between students in the same way as the former.[1] Nevertheless, the incorporation in examinations of more realistic, application-orientated tasks would seem to be desirable, not only for the sake of appearances, but also in the hope of encouraging the learning of more relevant facts and skills.

8.4 A critique of the methodology employed and some suggestions for further research

Two of the research techniques employed in the study, namely protocol analysis and associative mapping, cannot yet be regarded as fully developed tools. It will therefore be appropriate, before discussing

---

1. This is confirmed, in the findings reported in Chapter 6, by the absence of interactions between performance and achievement tertiles.

suggestions for further research, to consider the potential of these methods.

8.4.1 **Methodological issues.** The collection of tape recorded protocols represents a unique way of obtaining direct information concerning cognitive aspects of PS behaviour. While the detailed approach to the collection of protocols will no doubt be developed further (see eg Pines et al, 1978; White 1979b), the simple "thinking aloud" technique used in the present study appeared to be well adapted to its purpose and to interfere minimally with the processes under observation. In situations where it may be practicable, the provision of some training for the subjects involved seems likely to be value. As regards the researcher, his role will vary somewhat depending on the demands of any particular study. Whilst it is clear that any intervention carries the danger of distorting the course of a subject's problem solution, nevertheless valuable information may be obtainable with minimal interference if the researcher is clear about his objectives, is working on the basis of a well developed theoretical framework and has training or experience in protocol collection. The analysis of protocol data, however, continues to present a challenge. The use of simple, predetermined categories (eg Hatley, 1979) ignores the richness of the data and has little to recommend it. Similarly dimensional interpretations (as in current work in Australia - see Section 5.2.2) have an a priori character and appear ill-equipped to deal with unexpected features in the data. The main strength of the thorough-going information processing approach, involving the use of problem behaviour graphs and the induction of production systems, appears to lie in computer simulations and, ultimately, in its potential to generate detailed and general descriptions of information states which appear to trigger particular patterns of processing. Where broader interpretations are of interest, the development of a simple category analysis through the process described in Section 4.2.1 as "iterative generalisation" seems likely to have most to offer. The principal difficulty associated with this method is the heavy demands which it makes on the researcher both as regards his time and his judgement. The use of network analysis (see Section 4.2.3) appears to offer some prospect of lessening these problems and bringing more discipline to the process (Bliss and Ogborn, 1979). With further inductive and analytical development, a network such as that presented in Figure 4.2 (after Mujib, 1980) could provide a framework for studies of PS in the context of science examination items (and, with appropriate modifications, in broader PS contexts). It can be argued that the semantic grammar represented by such a network would help to maintain the stability of meaning of the strategy and error types identified and would provide an agreed framework against which particular instances and ambiguities might be interpreted.

The use of word association methods as a basis for cognitive mapping was reviewed in Section 5.2.5 The derivation of relatedness matrices from lists of continued associations using Garskoff and Houston's (1963) method of analysis can be regarded as well established, and the conversion of such matrices into group cognitive maps using Waern's (1972) multi-step graphical method has found increasing application in science. The striking similarity between the mean cognitive map generated in this way during the present study and the structure of the corresponding subject matter, echoes a number of recent similar findings. The possibility of mapping cognitive structures has been seen as opening the way for the testing of certain learning theories, in particular that of Ausubel, and for the development of their explanatory and predictive power (Kempa, 1982; see also Ausubel et al, 1978). The maps developed in the course of the present study appear promising in this respect, particularly in relation to the systematic differences between the maps of high and low achievers. However, there would seem to be scope both for improving cognitive mapping techniques and for the explicit evaluation of their explanatory and predictive power. While techniques other than those involving word association may have a part to play, discussion here will be limited to that method, the advantages of which in regard to its minimal interference with the structure which is being mapped and its ease of administration, have been discussed at length in Chapter 5 (Sections 5.2 and 5.4).

As regards the collection of data, it was apparent from the present study that only "constrained" associations (that is to say, associations of words in their scientific context) were of much value, and instructions to subjects should therefore be explicit in this respect. It was also clear that the employment of a shortened response time of 30 seconds helped to reduce failures of set. However, a small number of responses did not appear to represent truly spontaneous associations and means of optimising the proportion of the latter would be worth considering. These might include the development of improved instructions to students and a further shortening of the time allowed for responding. As regards the latter it was clear, in interpreting the relatedness thresholds employed in generating the cognitive maps, that the structure of a map was largely determined by a student's first two or three responses to each stimulus word.

A second consideration as regards the improvement of data collection, concerns the importance of covering an adequate list of concepts within any particular topic area. The adequacy of such a list can only be judged in the context of a particular research design, however it may be noted that, in the present study, topic coverage that was adequate for the general interpretation of group maps was often not so for more detailed work with individual maps. The employment of a still shorter response time, as suggested above, would facilitate the use of longer lists of concepts.

As regards the quality of the maps derived from word association data, improved response spontaneity and extended concept lists should lead to improved definition in the maps and thus to enhanced interpretive power. However, the selection of relatedness thresholds and the interpretation of the different levels of relatedness (inter-concept assessibility) represented by each, may also be areas in which development is possible. In particular, it is the author's impression that direct reference back to raw data from the response lists might greatly assist in the interpretation of **individual** maps. Indeed it might be of interest to evaluate an alternative method of constructing individual maps directly from the raw data using directional links after the manner employed by Cochaud (1979). Such a method might employ a number of a subject's most spontaneous associations (up to a maximum of, say, three per stimulus word) in place of Cochaud's single responses, resulting in a more complex, but possibly a more powerful, map.

8.4.2 **Suggestions for further research.** The broad and exploratory nature of the present study has been emphasised and its implications for further research are correspondingly wide. It should be sufficient, however, to indicate briefly below five main areas within which detailed research may be of interest.

Two possible fields for further investigation represent straightforward developments of the present work. Thus it would be of interest to refine the existing analysis of students' PS strategies and errors, possibly using a network approach to the analysis of protocol data, and to expand it to cover mathematical and objective test items; such analyses might also, of course, be extended to non-examination contexts and to other subject areas. Secondly, it would be of interest to replicate the word association study using improved mapping techniques in conjunction with the collection and analysis of protocols. The latter would, in particular, enable the interpretive power of individual maps to be assessed with more confidence. Indeed, comparisons between maps and detailed records of students' attempts might well contribute to the development of both mapping and protocol techniques. More fundamentally, such a study appears to offer the prospect, referred to earlier, of contributing to the development of a learning theory which has explanatory and predictive power.

It is in the last mentioned connection that a third area of research may be indicated. This would involve a development of experimental work of the kind reported by Mayer and Greeno (1972, 1974) and by Mayer (1975). In these studies, different teaching approaches were found to be linked to different outcomes in terms of LTM organisation, the latter being inferred by subsequent performance on various post-test tasks. The addition of cognitive mapping techniques to similar studies would offer the prospect of direct confirmation of the nature of the semantic structures resulting from particular teaching approaches, and of the relationship of such structures to subsequent cognitive performances. Such an approach appears to offer the exciting prospect of an empirical evaluation of alternative theories of learning. At least one preliminary report of work along these lines, using modified word association techniques, has already appeared (Gunstone, 1981).

A fourth area which might be of interest is suggested by work reported by Cochaud and Thompson (1979) concerning the use of single response word association methods by science teachers as a basis for planning and evaluating their teaching (see also Cochaud, 1979). Further research and development in relation of this and more sophisticated cognitive mapping techniques might be of value for monitoring the progress of learning and identifying cases where serious deficiencies in semantic networks may indicate the need for remedial teaching. The early identification and correction of such "error factors" could do much to improve the quality of the learning process.

Finally, there appears to have been little previous research into the influence of different levels of information on PS behaviour. The present exploratory study was therefore breaking new ground. In particular, certain items (which also had a role to play in other quite separate aspects of the study) did not lend themselves particularly well to being rewritten so as to incorporate different levels of information. In view of the interest of the findings as regards the relationship between PS performance in examinations and in more realistic situations, this appears to be a promising area for more detailed investigation. A study using tailor-made, rather than "general purpose", items and possibly supported by both cognitive mapping and protocol analysis, would provide an opportunity to confirm and extend the rather tentative findings which have been reported here.

8.5 Conclusion

The aim of the study, in the broadest sense, was to find out more about what it is that examinations measure as the end product of secondary school chemistry studies. The results suggest that examination success depends largely on the organisation of chemical concepts in a student's LTM, that is to say on the quality of the relevant parts of his semantic network. The study also indicates that the majority of O-level candidates may have difficulty in evaluating the demands of a chemical problem in terms of the information needed to solve it, and may thus be ill-equipped to tackle such problems in more realistic situations beyond the examination hall. Moreover, it seems probable that this difficulty, in turn, may reflect defects in a student's semantic network. On the basis of such conclusions, an important task for further research will be to identify or develop approaches to the teaching of science which optimise the quality of the semantic networks acquired by the learner.

REFERENCES

Andersson B. 1979, Some Aspects of Children's Understanding of Boiling Point, in Archenhold W.F., Driver R.H., Orton A. and Wood-Robinson C. (Eds), Cognitive Development Research in Science and Mathematics, (University of Leeds).

Archenhold W.F., Driver R.H., Orton A. and Wood-Robinson C. 1979, Cognitive Development Research in Science and Mathematics, (University of Leeds).

Ashmore A.D. 1976, Examinations and Problem-Solving in Degree Level Chemistry, unpublished PhD thesis (University of East Anglia; Norwich).

Ashmore A.D., Frazer M.J. and Casey R.J. 1979, Problem-Solving and Problem-Solving Networks in Chemistry, Journal of Chemical Education, Vol 56, pp 377-379.

Ausubel D.P. with Novak J.D. and Hanesian H. 1978, Educational Psychology: A Cognitive View (Holt, Rinehart and Winston; New York).

Ausubel D.P. and Robinson F.G. 1971, School Learning: Introduction to Psychology (Rinehart and Winston; New York).

Atkin J.A 1978, Cognitive Functioning: A Model for Learning and Problem-Solving, paper presented at the annual meeting of the National Association for Research in Science Teaching (Toronto, Canada).

Baddeley A.D. 1976, The Psychology of Memory (Harper and Row; London).

Bapat J.B. 1975, An Attempt to Develop a Methodology of Studying Problem-Solving Strategies, unpublished MSc thesis (University of East Anglia; Norwich).

Bartlett F.C. 1932, Remembering: A Study in Experimental and Social Psychology (Cambridge University Press; London).

Belanger M. 1969, Learning Studies in Science Education, Review of Educational Research, Vol 39, pp 377-395.

Bhaskar R. and Simon H.A. 1977, Problem-Solving in Semantically Rich Domains: Example from Engineering Thermodynamics, Cognitive Science, Vol 1, pp 193-215.

Bliss J. and Ogborn J. 1979, The Analysis of Qualitative Data, European Journal of Science Education, Vol 1, pp 427-440.

Bloom B.S. (Ed) 1956, Taxonomy of Educational Objectives - Book 1, Cognitive Domain (David McKay; New York).

Bloom B.S. and Broder L.J. 1950, Problem-Solving Processes in College Students, Supplementary Educational Monographs, No 73, pp 1-31 (University of Chicago Press), abridged in Harris T.L. and Schwahn W.E. (Eds) 1961, Selected Readings on the Learning Process (Oxford University Press; New York).

Bruner J.S. 1963, The Process of Education (Harvard University Press; Cambridge, Mass).

Bruner J.S., Goodnow J.J. and Austin G.A. 1956, A Study of Thinking (John Wiley; New York).

Butcher H.J. 1970, Human Intelligence (Methuen; London).

Cachapuz A.F.C. 1979, An Investigation into Relationships between Word Meaning and Achievement in the Learning of Reaction Kinetics, unpublished MSc thesis (University of East Anglia; Norwich).

Carlson J.S. et al 1977, A Comparison of the Predictive Validity of a Measure of General Intelligence and a Piaget-derived Test Relative to an Achievement Examination in High School Chemistry, Educational and Psychological Measurement, Vol 37, pp 999-1003 (quoted in ERIC abstract EJ 180486 TM 503158).

Cassels J.R.T. and Johnstone A.H. 1977, Language in Chemistry, in Research for the Classroom and Beyond (The Chemical Society; London).

Cassels J.R.T. and Johnstone A.H. 1980, The Understanding of Non-Technical Words in Science (The Chemical Society, London).

Cassels J.R.T. and Johnstone A.H. 1983, The Meaning of Words and the Teaching of Chemistry, Education in Chemistry, Vol 20, pp 10-11.

Champagne A.B. and Klopfer L.E. 1977, A Sixty-Year Perspective on Three Issues in Science Education: I Whose Ideas are Dominant? II Representation of Women III Reflective Thinking and Problem-Solving, Science Education, Vol 61, pp 431-452.

Champagne A.B., Klopfer L.E., Desena A.T. and Squires D.A. 1981, Structural Representations of Students' Knowledge before and after Science Instruction, Journal of Research in Science Teaching, Vol 18, pp 97-111.

Chao C.C. 1974, A Study of Conceptual Elements Involved in Two Physics Terms for Students of Different Cultural Backgrounds, unpublished PhD thesis (Ohio State University: quoted in ERIC abstract ED 093684 SE 018000).

Chokotho N.C.J. 1975, An Investigation into the Validity of the Properties of Bloom's Taxonomy, unpublished MSc thesis (University of East Anglia; Norwich).

Chornayko D.M., Christmas R.J., Cosic S., Dibbs S.E., Hamielec C.M., Macloed L.K., Moore R.F., Norman S.L., Stankowich R.J., Tyne S.C., Wong L.K. and Woods D.R. 1979, What is Problem-Solving?, Chemical Engineering Education, Vol 13, pp 132-137.

Cochaud G.A.M. 1979, Rules for Sequencing Material to be Learnt in Science, unpublished PhD Thesis (University of Oxford).

Cochaud G.A.M. and Thompson J.J. 1979, Sequencing Science Material for particular Classes, in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

Cofer C.N. 1958, Comparisons of Word Associations obtained by the Method of Discrete Single Word and Continued Association, Psychological Reports, Vol 3, pp 507-510.

Cowan J. 1977, Individual Approaches to Problem-Solving, paper presented at APLET conference (Department of Civil Engineering: Learning Unit, Heriot Watt University; Edinburgh).

Cramer P. 1966, Mediated Priming of Associative Responses: The Effect of Time-Lapse and Interpolated Activity, Journal of Verbal Learning and Verbal Behaviour, Vol 5, pp 163-166.

Cramer P. 1968, Word Association (Academic Press; New York: quoted in Preece 1978a op cit).

Curragh E.F. and Savage N.T. 1975, Attainment in Chemistry and its Relationship with other Cognitive Abilities, in Kempa R.F., Hoare D.E. and Rutherford R.J.D. (Eds), Research in Assessment (The Chemical Society; London).

Dangerfield A.S. 1977, An Investigation of Whether Specifically Written Materials can Influence Cognitive Structure, unpublished MSc thesis (University of East Anglia; Norwich)

de Bono E. 1967, The Five-Day Course in Thinking (Penguin; London).

de Bono E. 1969, The Mechanism of Mind (Jonathan Cape; London).

Deese J.E. 1962, On the Structure of Associative Memory, Psychological Review, Vol 69, pp 161-175.

Deese J.E. 1965, The Structure of Associations in Language and Thought (John Hopkins Press; Baltimore).

Dewey J. 1910, How We Think (D.C.Heath; Boston).

Dietze A.G. 1963, Types of Emotions or Dimensions of Emotions? A Comparison of Typal Analysis with Factor Analysis, Journal of Educational Psychology, Vol 56, pp 143-159.

Driver R. 1978, When is a Stage not a Stage?, Educational Research, Vol 21, pp 54-61.

Driver R. 1981, Pupils' Alternative Frameworks in Science, European Journal of Science Education, Vol 3, pp 93-101.

Driver R. 1983, The Pupil as Scientist? (Open University Press; Milton Keynes).

Dunker K. 1945, On Problem-Solving, Psychological Monographs Vol 58, No 5, whole No 270, extract abridged in Harris T.L. and Schwahn W.E. (Eds) 1961, Selected Readings on the Learning Process (Oxford University Press; New York).

Durkin H.E. 1937, Trail-and-Error, Gradual Analysis and Sudden Reorganisation: An Experimental Study of Problem-Solving, Archives of Psychology, Vol 30, No 210 abridged in Harris T.L. and Schwahn W.E. (Eds) 1961, Selected Readings on the Learning Process (Oxford University Press; New York).

Ebel R.L. 1965, Measuring Educational Achievement (Prentice Hall; Englewood Cliffs, N.J.).

Eggleston J.F. and Kerr J.F. 1968, Studies in Assessment (English Universities Press; London).

Everitt B.S. 1974, Cluster Analysis (Heinemann Educational Books; London).

Fenker R.M. 1975, The Organisation of Conceptual Materials: A Methodology for Measuring Ideal and Actual Structures, Instructional Science, Vol 4, pp 33-57.

Fisher R.A. 1979, Probing Cognitive Structure (School of Education, University of Leicester).

Flaherty E.G. 1975, The Thinking Aloud Technique and Problem-Solving Ability, Journal of Educational Research, Vol 68, pp 223-225.

Gable D.L. and Sherwood R.D. 1983, Facilitating Problem-Solving in High School Chemistry, Journal of Research in Science Teaching, Vol 20, pp 163-177.

Gagné R.M. 1966, Human Problem Solving: Internal and External Events, in Kleinmuntz B. (Ed), Problem-Solving Research, Method and Theory (Wiley; New York).

Gagne R.M. 1970, The Conditions of Learning (Holt, Rinehart and Winston; New York).

Gagné R.M. and Smith E.C. 1962, A Study of the Effects of Verbalisation on Problem-Solving, Journal of Experimental Mycology, Vol 63, pp 12-18.

Gagné R.M. and White R.T. 1978, Memory Structure and Learning Outcome, Reviews of Educational Research, Vol 48, pp 187-222.

Gardner P.L. 1972, Words in Science (Australian Science Education Project; Melbourne).

Gardner R.W., Holzman P.S., Klein G.S., Linton H.B. and Spence D.P., 1959, Cognitive Control: A Study of Individual Consistencies in Cognitive Behaviour, Psychological Issues, Vol 1, No 4.

Garskoff B.E. and Houston J.P. 1963, Measurement of Verbal Relatedness: An Ideographical Approach, Psychological Reviews, Vol 70, pp 277-288.

Gilbert J.K. and Osborne R.J. 1979, Identifying Science Students' Concepts: the Interview-About-Instances Approach in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

Greene J. 1972, Psycholinguistics: Chomsky and Psychology (Penguin Education; London).

Greeno J.G. 1973, The Structure of Memory and the Process of Solving Problems in Soloso R.L. (Ed) Contemporary Issues in Cognitive Psychology - the Loyola Symposium (Winston; Washington D.C.).

Greeno J.G. 1976, Cognitive Objectives of Instruction: Theory of Knowledge for Solving Problems and Answering Questions in Klohr D. (Ed), Cognitive and Instruction (Erlbaum Associates; Hillsdale, N.J.).

Greeno J.G. 1978, Nature of Problem-Solving Abilities, in Estes W.K. (Ed), Handbook of Learning and Cognitive Processes - Vol 5 (Lawrence Erlbaum Associates; Hillsdale N.J.).

Greeno J.G. 1980a, Analysis of Understanding in Problem-Solving in Kleuse R.H. and Spada H. (Eds), Developmental Models of Thinking (Academic Press; New York).

Greeno J.G. 1980b, Trends in the Theory of Knowledge for Problem-Solving, in Tuma D.T. and Reif F. (Eds), Problem-Solving and Education (Lawrence Erlbaum Associates; Hillsdale N.J.).

Gregg L.W. and Farnham-Diggory S. 1975, Content and Structure in Learning (National Institute of Mental Health; Rockville Md).

Harlen W., Black P. and Johnson S. 1981, Science in Schools: Age 11 Report No 1 (Her Majesty's Stationery Office; London).

Harlow H.F. 1949, The Formation of Learning Sets, Psychological Review, Vol 56, pp 51-65.

Harrison A. 1982, Review of Graded Tasks, Schools Council Examinations Bulletin, 41 (Methuen Educational; London).

Hateley R.J. 1979, Answering Fixed Response Items in Chemistry - A Pilot Study, School Science Review, Vol 60, No 212, pp 568-570 .

Hoare D.E., Kempa R.F. and Ongley P.A. (undated C 1978), Research in Assessment II (The Chemical Society, London).

Hollander S.K. 1978, Thought Processes Employed in the Solution of Verbal Arithmetic Problems, School Science and Mathematics, Vol 78, pp 327-334.

Holzman P.S. and Gardner R.W. 1972, Levelling-Sharpening and Memory Organisation in L. Sperry (Ed), Learning Performance and Individual Differences (Scott Foresman; Illinois).

Inhelder B. and Piaget J. 1958, The Growth of Logical Thinking from Childhood to Adolescence, translated by Parsons A. and Milgram S. (Routledge and Kegan Paul; London).

Isa A.M. and Maskill R. 1982, A comparison of Science Word Meaning in the Classrooms of Two Different Countries: Scottish Integrated Science in Scotland and Malaysia, British Journal of Educational Psychology, Vol 52, pp 188-198.

Jackson K.F. 1975, The Art of Solving Problems (Heinemann; London).

Jardine N. and Sibson, R. 1968, The Construction of Hierarchic and Non-hierarchic Classifications, The Computer Journal, Vol 11, pp 177-184.

Johnson C.J. 1976, The Investigation of Cognitive Structure by Directed Graph Analysis of Essay Materials, unpublished MSc thesis (University of East Anglia; Norwich).

Johnson P.E. 1964, Associative Meaning of Concepts in Physics, Journal of Educational Psychology, Vol 55, pp 84-88.

Johnson P.E. 1965, Word Relatedness and Problem-Solving in High School Physics, Journal of Educational Psychology, Vol 56, pp 217-224.

Johnson P.E. 1967, Some Psychological Aspects of Subject Matter Structure, Journal of Educational Psychology, Vol 58, pp 75-83.

Johnson P.E. 1969, On the Communication of Concepts in Science, Journal of Educational Psychology, Vol 60, pp 32-40.

Johnson P.E. 1971, Perceptual Structure of Scientific Knowledge (Office of Education, Bureau of Research, US Department of Health, Welfare and Education; Washington DC).

Johnson P.E., Cox D.L. and Curran T.E. 1970, Psychological Reality of Physical Concepts, Psychonomic Science, Vol 19, pp 245-246.

Johnson P.E., Cox D.L. and Curran T.E. 1971, A Model for Knowledge of Science Concepts, Journal of Research in Science Teaching, Vol 8, pp 91-95.

Johnson S.C. 1967, Hierarchical Clustering Schemes, Psychometrika, Vol 32, pp 241-254.

Johnstone A.H. 1981, Chemical Education Research - Facts, Findings and Consequences, Chemistry in Britain, Vol 17, pp 130-135.

Johnstone A.H. and Kellett N.C. 1980, Learning Difficulties in School Science: Towards a Working Hypothesis, European Journal of Science Education, Vol 2, pp 175-181.

Johnstone A.H. and Letton K.M. 1982, Recognising Functional Groups, Education in Chemistry, Vol 19, pp 16-19.

Joint Matriculation Board 1979, General Certificate of Education: Regulations and Syllabuses (Joint Matriculation Board; Manchester).

Joseph J. 1974, Problem-Solving is Everybody's Problem, Science Teacher, Vol 41, pp 16-18.

Kagan J. 1966, Reflection-Impulsivity: The General Dynamics of Conceptual Tempo, Journal of Abnormal Psychology, Vol 71, pp 17-24.

Kagan J. 1972, Reflection-Impulsivity: The Generality and Dynamics of Conceptual Tempo, in Sperry, L. (Ed), Learning Performance and Individual Differences (Scott, Foresman; Illinois).

Kass H. 1971, Structure in Perceived Relations Among Physics Concepts, Journal of Research in Science Teaching, Vol 8, pp 339-350.

Kempa R.F. 1979, The Future, in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

Kempa R.F. 1982, Learning Theories and Chemical Education in Chemical Education Research - Implications for Teaching (Royal Society of Chemistry; London).

Kempa R.F., Hoare D.E. and Rutherford R.J.D. 1975, Research in Assessment (The Chemical Society; London).

Kempa R.F. and L'Odiaga J. 1983, Criterion-Referenced Interpretation of Norm-Referenced Examination Grades: A Feasibility Study in Ongley P.A., Brockington J. and Sheridan P. (Eds), Research in Assessment III (The Royal Society of Chemistry; London).

Kempa R.F. and Nicholls C.E. 1983, Problem-Solving Ability and Cognitive Structure: An Exploratory Investigation, European Journal of Science Education, Vol 5, pp 171-184.

Keohane K.W. 1977, Research on Science Education in the United Kingdom, in Frey K., Blansdorf K., Kapune Th. and Schaefer G. (Eds), Research on Science Education in Europe (Lisse; Netherlands).

Kerlinger F.N. 1964, Foundations of Educational Research (Holt, Rinehart and Winston; London).

Khan A.A. 1980, Problem-Solving Difficulties in Electrophilic Aromatic Substitution and the Synthesis of Polysubstituted Benzenes, unpublished MSc thesis (University of East Anglia; Norwich).

Klein G.A. and Weitzenfeld J. 1976, General Description of Human Problem-Solving (Air Force Human Resources Laboratories, Advanced System Division; Brooks AFB, Texas and Wright-Patterson AFB, Ohio).

Kruskal J.B. 1964, Multidimensional Scaling by Optimising Goodness of Fit to a Nonometric Hypothesis, Psychometrika, Vol 29, pp 1-27.

Krutetskii V.A. 1976, The Psychology of Mathematical Abilities in School Children, translated by Teller J. and edited by Kilpatrick J. and Wirszup I. (University of Chicago Press; London).

Laffal J. and Feldman S. 1962, The Structure of Single Word and Continuous Associations, Journal of Verbal Learning and Verbal Behaviour, Vol 1, pp 54-61.

Larkin J.H. 1980, Teaching Problem-Solving in Physics, in Tuma D.T. and Reif F. (Eds), Problem-Solving in Education (Lawrence Erlbaum Associates; Hillsdale N.J.).

Larkin J.H. and Reif F. 1979, Understanding and Teaching Problem-Solving in Physics, European Journal of Science Education, Vol 1, pp 191-203.

Lewis D.G. 1961, Group Factors in Attainment in General Science Subjects, British Journal of Educational Psychology, Vol 31, pp 241-248.

Lewis D.G. 1964, The Factorial Nature of Attainment in Elementary Science, British Journal of Educational Psychology, Vol 34, pp 1-9.

Lewis D.G. 1969a, Ability in Science at the Ordinary Level of the General Certificate of Education, British Journal of Educational Psychology, Vol 37, pp 361-370.

Lewis D.G. 1969b, Categories of Ability in Elementary Science, Journal of Research in Science Teaching, Vol 6, pp 143-149.

Lewis D.G. 1973, Statistical Methods in Education (University of London Press; London).

Lindsay P.H. and Norman D.A. 1977, Human Information Processing (Academic Press; New York).

L'Odiaga J. 1977, Feasibility Study of Criterion Referenced Testing in Advanced-level Chemistry, unpublished MSc thesis (University of Keele).

MacGuire P.R.P. 1982, Techniques for Investigating the Understanding of Concepts in Science, in Chemical Educational Research - Implications for Teaching (The Royal Society of Chemistry; London).

Maier N.R.F. 1930, Reasoning in Humans: I On Direction, Journal of Comparative Psychology Vol 10, pp 115-143 abridged in Harris T.L. and Schwahn W.E. (Eds) 1961, Selected Readings on the Learning Process (Oxford University Press; New York).

Malin J.T. 1979, Strategies in Mathematical Problem-Solving, Journal of Educational Research, Vol 73, pp 101-108.

Mallison G.G. 1969, Factors Affecting College Students' Achievement in Science: Final Report (West Michigan University; Kalamazoo as quoted in ERIC abstract ED 028100 24 SE 006547).

Mandell A. 1980, Problem-Solving Strategies of 6th-grade Students who are Superior Problem-Solvers, Science Education, Vol 64 pp 203-211.

Massey A.J. 1982, Assessing 16+ Chemistry: The Exposure-Mastery Gap, Education in Chemistry, Vol 19, pp 143-145.

Mayer R.E. 1975, Information Processing Variables in Learning to Solve Problems, Reviews of Educational Research, Vol 45, pp 525-541.

Mayer R.E. 1977, Thinking and Problem-Solving (Scott-Foresman; Glenviers, Illinois).

Mayer R.E. and Greeno J.G. 1972, Structural Differences Between Learning Outcomes Produced by Different Instructional Methods, Journal of Educational Psychology, Vol 63, pp 165-173.

Mayer R.E. and Greeno J.G. 1974, Restructuring of Meaningful Information in Problem-Solving, ERIC Research Reports (reference ED 108961 - SE 019462).

Maruping M. 1980, The Measurement of Word Meaning and its Relationships to Learning, unpublished MSc thesis (University of East Anglia; Norwich).

Maskill R. and Pereira D. 1979, Cognitive Structure from Digraph Analysis in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

McCabe R.E. 1975, A Study of Students' Difficulties in Chemical Problem-Solving, unpublished MSc thesis (University of East Anglia; Norwich).

McQuitty L.L. 1966, Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data, Educational and Psychological Measurement, Vol 26, pp 825-831.

Mead M.A. 1972, RCMAT - A Fortran Program to Calculate Relatedness Coefficients among Stimulus Words by Comparing Lists of Verbal Responses to Stimuli, Behavioural Science, Vol 17, p 568 (quoted in Cachapuz 1979, op cit).

Michon J.A. 1972, Multidimensional and Hierarchical Analysis of Progress in Learning, in Gregg L.W. (Ed) Cognition in Learning and Memory (Wiley; New York).

Miller G.A. 1956, The Magic Number 7 Plus or Minus 2, Psychological Review, Vol 63, pp 81-97.

Moreira M.A. and Santos C.A. 1981, The Influence of Content Organisation on Student's Cognitive Structure in Thermodynamics, Journal of Research in Science Teaching, Vol 18,. pp 525-531.

Morris J. 1977, The Development and use of an Analytical Model for the Investigation of Problems and Problem-Solving in Stoichiometry, unpublished MSc thesis (University of East Anglia; Norwich).

Morris V.D. 1970, An Analysis of Changes in Associative Responses to Selected Science Concept Words by Seventh-Grade Pupils During Study in a Modern Science Curriculum, unpublished PhD thesis (Florida State University).

Moynihan T.F. 1982, Word Association Techniques for the Study of Cognitive Structure of Chemical Concepts in Scottish Secondary Schools, in Chemical Education Research - Implications for Teaching (The Royal Society of Chemistry; London).

Mujib F. 1980, Students' Problem-Solving Behaviours in Selected Problems Related to Undergraduate Physics, with Specific Attention to the Analysis of Protocols, unpublished PhD thesis (University of London).

Nagy P. 1983, Assessing Cognitive Structure: A Response to Stewart, Science Education, Vol 67, pp 25-36.

Neie Van E. 1972, Verbal Predictive Ability and Performance on Selected Science Process Tasks, Journal of Research in Science Teaching, Vol 9, pp 213-221.

Newell A. 1977, On the Analysis of Human Problem-Solving Protocols in Johnson-Laird P.N. and Wason P. (Eds) Thinking: Readings in Cognitive Science (Cambridge University Press; London).

Noble C.E. 1952, An Analysis of Meaning, Psychological Review, Vol 52, pp 421-430.

Norman D.A. 1973, Memory, Knowledge and the Answering of Questions, in Soloso R.L. (Ed), Contemporary Issues in Cogntive Psychology; the Loyola Symposium (Winston; Washington D.C.).

Novak J.D. 1976, Understanding the Learning Process and Effectiveness of Teaching Methods in the Classroom, Laboratory and Field, Science Education, Vol 60, pp 493-506.

Novak J.D. 1978, An Alternative to Piagetian Psychology for Science and Mathematics Education, Studies in Science Education, Vol 5, pp 1-30.

Noval L.K. 1976, A Computer Program for Deriving Measures of Verbal Relatedness, Educational and Psychological Measurement, Vol 36, pp 217-221 (quoted in Cachapuz 1979, op cit).

Nussbaum J. and Novick S. 1981, Brainstorming in the Classroom to Invent a Model - A Case Study, School Science Review, Vol 62, pp 771-778.

Ogborn J. 1979, Some Uses of Networks of Options for Describing Complicated Qualitative Data, in Archenhold W.F., Driver R.H., Orton A. and Wood-Robinson C. (Eds), Cognitive Development Research in Science and Mathematics (The Centre for Studies in Science Education; University of Leeds).

Omerod M.B. and Duckworth D. 1975, Pupils' Attitudes to Science (NFER Publishing Company; Windsor).

Ongley P.A., Brockington J. and Sheridan P. 1983, Research in Assessment III (Royal Society of Chemistry; London).

Osborne R.J. 1980, A Method of Investigating Concept Understanding in Science, European Journal of Science Education, Vol 2, pp 311-321.

Pereira D. 1976, The Structural Analysis of Instructional Material Using Di-graph Techniques, unpublished MSc thesis (University of East Anglia; Norwich).

Pereira D. 1979, The Structure of Communication and the Learning of Chemistry, unpublished PhD thesis (University of East Anglia; Norwich).

Pereira D.J.V.C. and Maskill R. 1983, Structure and Process in Pupils' Essays: A Graphical Analysis of the Organisation of Extended Prose, British Journal of Educational Psychology, Vol 53, pp 100-106.

Pickthorne B. 1983, Error Factors: A Missing Link Between Cognitive Science and Classroom Practice, Instructional Science, Vol 1, pp 281-291.

Pines A.L. 1977, Scientific Concept Learning in Children: The Effect of Prior Knowledge on Resulting Cognitive Structure Subsequent to A-T Instruction, unpublished PhD thesis (Cornell University; New York).

Pines A.L. 1980, Protocols as Indicators of Cognitive Structure: A Cautionary Note, Journal of Research in Science Teaching, Vol 17, pp 361-362.

Pines A.L., Novak J.D., Posner G.J. and Van Kirk J. 1978, The Clinical Interview: A Method for Evaluating Cognitive Structure, unpublished report (Department of Education, Cornell University; New York).

Polya G. 1957, How to Solve It (Doubleday Anchor Books; New York).

Polya G. 1968, Mathematical Discovery, Volume II: On Understanding Learning and Teaching Problem-Solving, (Wiley; New York).

Posner G.J. and Gertzog W.A. 1982, The Clinical Interview and the Measurement of Conceptual Change, Science Education, Vol 66, pp 195-209.

Powell A.V. 1976, A Study of the Abilities Required for Chemical Problem-Solving, unpublished MSc thesis (University of East Anglia; Norwich).

Preece P.F.W. 1976a, Mapping Cognitive Structure: A Comparison of Methods, Journal of Educational Psychology, Vol 68, pp 1-8.

Preece P.F.W. 1976b, Associative Structure of Science Concepts, British Journal of Educational Psychology, Vol 46, pp 174-183.

Preece P.F.W. 1976c, The Concepts of Electromagnetism: A Study of the Internal Representation of an External Structure, Journal of Research in Science Teaching, Vol 13, pp 517-524.

Preece P.F.W. 1976d, Science Concepts in Semantic Space - A Multi-dimensional Scaling Study, Alberta Journal of Educational Research, Vol 22, pp 281-288.

Preece P.F.W. 1977, Development Trends in Continued Word Associations of Physics Students, Journal of Research in Science Teaching, Vol 14, pp 235-239.

Preece P.F.W. 1978a, Associative Structure and The Schema of Proportionality, Journal of Research in Science Teaching, Vol 15, pp 395-399.

Preece P.F.W. 1978b, Exploration of Semantic Space: Review of Research on The Organisation of Scientific Concepts in Semantic Memory, Science Education, Vol 62, pp 547-562.

Reif F., Larkin B.H. and Brackett G.C. 1976, Teaching General Learning and Problem-Solving Skills, American Journal of Physics, Vol 44, pp 212-217.

Reitmann W.R. 1965, Cognition and Thought: An Information Processing Approach (Wiley; New York).

Ring D.G. and Novak J.D. 1971, The Effects of Cognitive Structure Variables on Achievement in College Chemistry, Journal of Research in Science Teaching, Vol 8, pp 325-333.

Rollins H.A. and Genser L. 1977, The Role of Cognitive Style in a Cognitive Task: A Case Favouring the Impulsive Approach to Problem-Solving, Journal of Educational Psychology, Vol 69, pp 281-287.

Ross K.A. and Sutton C.R. 1982, Concept Profiles and the Cultural Context, European Journal of Science Education, Vol 4, pp 311-323.

Rothkoph E.Z. and Thurner R.D. 1970, Effects of Written Instructional Material on the Statistical Nature of Test Essays, Journal of Educational Psychology, Vol 61, pp 83-89.

Rowell R.M. 1975, An Investigation of Children's Concepts of Natural Phenomena, paper presented at the Annual Conference of the National Association for Research in Science Teaching (Los Angeles).

Rumelhat D.E., Lindsay P.H. and Norman D.A. 1972, A Process Model for Long-Term Memory, in Tulving E. and Donaldson W. (Eds), Organisation of Memory (Academic Press; New York).

Schaefer G. 1979a, Concept Formation in Biology: The Concept of Growth, European Journal of Science Education, Vol 1, pp 87-101.

Schaefer G. 1979b, Inclusive Thinking with Inclusive Concepts, in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

Selvaratnam M. 1983, Students' Mistakes in Problem-Solving, Education in Chemistry, Vol 20, pp 125-132.

Selvaratnam M. and Frazer M.J. 1982, Problem-Solving in Chemistry (Heinemann Educational Books; London).

Shann M.H. 1976, Measuring Problem-Solving Skills and Processes in Elementary School Children (School of Education, Boston University).

Shavelson R.J. 1972, Some Aspects of the Correspondence Between Content and Cognitive Structure in Physics Instruction, Journal of Educational Psycholgy, Vol 63, pp 225-234.

Shavelson R.J. 1973, Learning from Physics Instruction, Journal of Research in Science Teaching, Vol 10, pp 101-111.

Shavelson R.J. 1974, Methods of Examining Representations of a Subject-Matter Structure in a Student's Memory, Journal of Research in Science Teaching, Vol 11, pp 231-249.

Shavelson R.J. and Geeslin W.E. 1975, A Method for Examining Subject Structure in Instructional Materials, Journal of Structural Learning, Vol 4, pp 199-218.

Shavelson R.J. and Lee H. 1975, RCMATDIF: A Fortran Program to Calculate Relatedness Coefficients on a Verbal Relatedness Task and to Compare the Mean Matrix of one Group of Students to Another using Hotelling's T^2, Behaviour Research Methods and Instruments, Vol 7, pp 311-312 (quoted in Cachapuz 1979, op cit).

Shavelson R.J. and Stanton G.C.1975, Construct Validation: Methodology and Application to Three Measures of Cognitive Structure, Journal of Educational Measurement, Vol 12, pp 67-85.

Shayer M. and Adey P. 1981, Towards a Science of Science Teaching (Heinemann Educational Press; London).

Shepard R.N. 1962, The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function, Psychometrika, Vol 27, pp 125-140.

Shulman L.S. and Tamir P. 1972, Research on Teaching in the National Sciences, in Travers R.M.W. (Ed), Second Handbook of Research on Teaching (Rand McNally; Chicago).

Silver E.A. 1979, Problem-Solving Performance and Memory for Mathematical Problems: Solving Related Problems, paper presented at the annual meeting of the American Educational Research Association (as quoted in ERIC abstract ED 175630 SE 027745).

Simon H.A. 1974, How Big is a Chunk?, Science, Vol 183, pp 482-488.

Simon H.A. and Newell A. 1972, Human Problem-Solving (Prentice Hall; Englewood Cliffs N.J.).

Simon D.P. and Simon H.A. 1978, Individual Differences in Solving Physics Problems, in Siegler R.S. (Ed) Children's Thinking: What Develops (Lawrence Erlbaum Associates; Hillsdale N.J.).

Slimming D. 1971, An Investigaton into some Aspects of Teacher-based Assessment of VIth form Chemistry Students' Interpretational Abilities, unpublished MSc thesis (University of East Anglia; Norwich).

Stewart J. 1979, Content and Cognitive Structure: A Critique of Assessment and Representation used by Science Education Researchers, Science Education, Vol 63, pp 395-405.

Stewart J. 1980, Techniques of Assessing and Representing Information in Cognitive Structure, Science Education, Vol 64, pp 223-235.

Sutton C.R. 1980, The Learner's Prior Knowledge: A Critical Review of Techniques for Probing its Organisation, European Journal of Science Education, Vol 2, pp 107-120.

Thorsland M.N. and Novak J.D. 1975, The Identification and Significance of Intuitive and Analytical Approaches among College Physics Students, Science Education, Vol 58, p 245-265.

Thro M.P. 1978, Relationships Between Associative and Content Structure of Physics Concepts, Journal of Educational Psychology, Vol 70, pp 971-978.

Vincent A. 1981, Volumetric Concepts - Student Difficulties, Education in Chemistry, Vol 18, pp 114-115.

Waern Y. 1972, Structure in Similarity Matrices: A Graphical Approach, Scandinavian Journal of Psychology, Vol 13, pp 5-16.

Wallace T.F. 1978, A Study of Problem-Solving Involving Inorganic Chemistry at VIth form level, unpublished MSc thesis (University of East Anglia; Norwich).

Ward J.E. 1982, The Interpretation and Interpretability of Observational Data of Different Levels of Quality by Students, in Chemical Education Research - Implications for Teaching (The Royal Society of Chemistry; London).

Wason P.C. and Johnson-Laird P.N. 1972, Psychology of Reasoning (Batsford; London).

Waterman D.A. and Newell A. 1971, Protocol Analysis as a Task for Artificial Intelligence, in Proceedings of Second International Joint Conference on Artificial Intelligence [Imperial College, London], (The British Computer Society; London).

Webb N.L. 1979, Processes, Conceptual Knowledge and Mathematical Problem-Solving Ability, Journal of Research in Mathematics Education Vol 10, pp 83-93.

Wertheimer M. 1959, Productive Thinking (Harper and Bros; New York).

West L.H.T. 1979, Towards Descriptions of the Cognitive Structures of Science Students, in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

West L.H.T., Fensham P.J. and Garrard I. 1982, Describing the Cognitive Structure of Undergraduate Chemistry Students, unpublished report (Faculty of Education, Monash University).

Weston A.J. 1983, The Munn and Dunning Reports Government Development Programme, Joint Working Party Report - Science (Scottish Education Department and Central Committee on the Curriculum; Edinburgh).

White R.T. 1979a, Describing Cognitive Structure, paper given at the Meeting of the Austrlian Association for Educational Research (Melbourne).

White R.T. 1979b, Interview Protocols and Dimensions of Cognitive Structure, unpublished report (Monash University).

White R.T. and Gunstone R.F. 1980, Converting Memory Protocols to Scores on Several Dimensions, paper given at the meeting of The Australian Association for Educational Research (Sydney).

Wierstra R.F.A 1979, Psychological Theories and Research Methods Relating to Pupils Mental Structures as Aids to Designing a Curriculum, in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

Wijenayake K.M.M. 1978, An Investigation into the Growth of Word Meaning in Science, unpublished MSc thesis (University of East Anglia; Norwich).

Stewart J. 1980, Techniques of Assessing and Representing Information in Cognitive Structure, Science Education, Vol 64, pp 223-235.

Sutton C.R. 1980, The Learner's Prior Knowledge: A Critical Review of Techniques for Probing its Organisation, European Journal of Science Education, Vol 2, pp 107-120.

Thorsland M.N. and Novak J.D. 1975, The Identification and Significance of Intuitive and Analytical Approaches among College Physics Students, Science Education, Vol 58, p 245-265.

Thro M.P. 1978, Relationships Between Associative and Content Structure of Physics Concepts, Journal of Educational Psychology, Vol 70, pp 971-978.

Vincent A. 1981, Volumetric Concepts - Student Difficulties, Education in Chemistry, Vol 18, pp 114-115.

Waern Y. 1972, Structure in Similarity Matrices: A Graphical Approach, Scandinavian Journal of Psychology, Vol 13, pp 5-16.

Wallace T.F. 1978, A Study of Problem-Solving Involving Inorganic Chemistry at VIth form level, unpublished MSc thesis (University of East Anglia; Norwich).

Ward J.E. 1982, The Interpretation and Interpretability of Observational Data of Different Levels of Quality by Students, in Chemical Education Research - Implications for Teaching (The Royal Society of Chemistry; London).

Wason P.C. and Johnson-Laird P.N. 1972, Psychology of Reasoning (Batsford; London).

Waterman D.A. and Newell A. 1971, Protocol Analysis as a Task for Artificial Intelligence, in Proceedings of Second International Joint Conference on Artificial Intelligence [Imperial College, London], (The British Computer Society; London).

Webb N.L. 1979, Processes, Conceptual Knowledge and Mathematical Problem-Solving Ability, Journal of Research in Mathematics Education Vol 10, pp 83-93.

Werthemier M. 1959, Productive Thinking (Harper and Bros; New York).

West L.H.T. 1979, Towards Descriptions of the Cognitive Structures of Science Students, in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

West L.H.T., Fensham P.J. and Garrard I. 1982, Describing the Cognitive Structure of Undergraduate Chemistry Students, unpublished report (Faculty of Education, Monash University).

Weston A.J. 1983, The Munn and Dunning Reports Government Development Programme, Joint Working Party Report - Science (Scottish Education Department and Central Committee on the Curriculum; Edinburgh).

White R.T. 1979a, Describing Cognitive Structure, paper given at the Meeting of the Austrlian Association for Educational Research (Melbourne).

White R.T. 1979b, Interview Protocols and Dimensions of Cognitive Structure, unpublished report (Monash University).

White R.T. and Gunstone R.F. 1980, Converting Memory Protocols to Scores on Several Dimensions, paper given at the meeting of The Australian Association for Educational Research (Sydney).

Wierstra R.F.A 1979, Psychological Theories and Research Methods Relating to Pupils Mental Structures as Aids to Designing a Curriculum, in Archenhold et al (op cit) Cognitive Development Research in Science and Mathematics (University of Leeds).

Wijenayake K.M.M. 1978, An Investigation into the Growth of Word Meaning in Science, unpublished MSc thesis (University of East Anglia; Norwich).

Wilby P. 1983, Question: Is the Exam Exciting?, The Sunday Times, 15th May, p 14.

Wishart D. 1979, CLUSTAN 1B (University of Manchester Regional Computing Centre).

Witkkin H.A., Moore C.A, Goodenough D.R. and Cox P.W. 1977, Field Dependent and Field Independent Cognitive Styles and their Educational Implications, Review of Educational Research, Vol 47, pp 6-44.

Woolley E. 1979, Saved by a Spot of Quick Thinking, The Guardian, 15th June, p 11.

Worsley C. 1979, Talk on the Work of the Assessment of Performance Unit delivered at the Science Teachers Centre, University of Keele.

Zerlinker T. and Jeffrey W.E. 1976, Reflective and Impulsive Children: Strategies of Information Processing Underlying Diffrences in Problem-Solving, Monographs in the Society for Research in Child Development, Vol 41, No 5 (University of Chicago Press).