Representation and Strategy in Reasoning: An Individual Differences Approach

Padraic Monaghan

PhD
The University of Edinburgh
2000
Abstract

Individual differences in reasoning have been observed in a wide variety of tasks. Descriptions of the variation in response have been framed in terms of use of different strategies that invoke different representations. This thesis argues that in order to convert descriptions into explanations of performance it is necessary to compare and combine psychometric accounts with computational accounts of the processes underlying representation selection and use. Descriptions of strategies, representations and algorithms and their inter-relationships are necessary for a full account of reasoning behaviour.

Two large-scale studies of deductive reasoning are presented to illustrate this approach in action, and the inadequacy of accounts that do not provide accounts at all these levels. The first compares two theoretically motivated methods for solving categorial syllogisms, the second study assesses learning from and learning within a multimodal logic course called Hyperproof. These studies are compared to measures of spatial ability, field-independence/dependence, and serialist/holist learning style. The interaction of students' styles of learning with different presentations of information generalises across the domains. This generality is best expressed when psychometric and computational accounts of reasoning are consolidated.
Declaration

I declare that this thesis has been composed by myself and that the research reported is my own, unless otherwise stated.

Padraic Monaghan
Acknowledgements

Many thanks to my supervisors Keith Stenning and Jon Oberlander whose complementary styles and enthusiasms maintained my own. Keith’s comment that “it only takes one zebra to make a species” has stayed with me, summing up the comfort and inspiration his supervision has had.

Also thanks to Richard Cox, John Lee and the other members of the Graphics and Language Working Group in the HCRC for their excellent advice and multiple representations. Richard Catrambone was a great host for a research trip at Georgia Tech in 1998 where this thesis began to take shape. Nils Dahlback and Max Roberts helped many of these ideas to crystallise. The ESRC provided a Research Studentship that supported this work.

Catherine Breslin, Scott McDonald, Richard Shillcock and Frances Wignall made sure everything remained in proportion. As did the Gentlemen and Players of Leith Franklin Academicals Cricket Club, who obdurately refused to be interested in the slightest in what I was working on.

This thesis is dedicated to my parents, who made it all possible long before I knew anything about it.
...it is our choices, Harry, that show what we truly are, far more than our abilities.

*Harry Potter and the Chamber of Secrets*, J.K. Rowling.

It was not logic that carried me on; as well one might say that the quicksilver in the barometer changes the weather. It is the concrete being that reasons; pass a number of years and I find myself in a new place; how? Paper logic is but the record of it.

*Apologia Pro Vita Sua*, John Newman
Contents

1 Individual differences and cognitive science
   1.1 The starting point ...................................... 1
   1.2 The psychometric approach ............................. 5
   1.3 Integrating individual differences ..................... 6
   1.4 Outline of the thesis .................................. 9

2 Strategies in complex problem solving .................... 11
   2.1 Strategies in reasoning ................................ 11
   2.2 Spatial and verbal: specificity and expressivity ........... 14
   2.3 Expressivity/specificity in reasoning ................... 16
      2.3.1 Linear Syllogisms ................................ 16
      2.3.2 Sentence-picture verification ...................... 22
      2.3.3 Compass Directions Task .......................... 26
   2.4 Spatial ability: what does it measure? .................. 30
   2.5 Summary of Chapter ................................... 34

3 Selecting tests and tasks ................................ 36
   3.1 Strategies and Styles .................................. 37
      3.1.1 Visual/verbal preference ......................... 37
# 6 Teaching methods for solving syllogisms

## 6.1 Method

6.1.1 Subjects

6.1.2 Materials and Procedure

6.1.3 Pre-tests

6.1.4 Teaching strategies

## 6.2 Results

6.2.1 Pre-tests

6.2.2 Performance on the syllogisms prior to instruction: spontaneous strategy development

6.2.3 Learning the methods for solving syllogisms

## 6.3 Discussion

## 6.4 Comparing different stages in the two methods

6.4.1 Translating-in stage

6.4.2 Manipulating stage

6.4.3 Translating-out stage

## 6.5 General discussion

## 6.6 Holes in the scholastic sieve

# 7 A continuum of problem types and developing strategies: Hyperproof II

## 7.1 A taxonomy of problem types

7.1.1 Problem types in the GRE

7.1.2 Generalising the taxonomy: problem types in the BW

7.1.3 Problem types and strategies in HP

## 7.2 Aims and expectations

## 7.3 Method
Chapter 1

Individual differences and cognitive science

1.1 The starting point

The role of representation in reasoning has been a central concern in studies of the psychology of problem solving. How is information represented during reasoning tasks, and why might certain types of presentation assist in reasoning? These questions have not yet been adequately answered, primarily because attention to individual differences has been lacking, or differences have been assumed to be at the wrong level of description.

Debates about the nature of the “fundamental reasoning mechanism” (FRM) – the engine that drives deductive reasoning processes – have concentrated on disagreements over the representations used by this mechanism. Rips (1994) has proposed that humans operate some form of “mental logic”, where information is represented as syntactic statements and reasoning proceeds by rules operating on these statements. In contrast, Erickson (1974) has suggested that the representations in the FRM are spatial, and operations are on sets of individuals or properties.

Another alternative was developed by Johnson-Laird and Steedman (1978), where the FRM is driven by mental models, or descriptions of individuals. Individual differences can be incorporated into all three of these accounts, provided the differences can be accommodated by variations in the parameters of each account. Therefore, individual differences will be qualitative,
in terms of how many individuals or operations can be supported by working memory, or how many alternative models can be constructed due to storage limitations. Individual differences were considered to revolve around different abilities for subjects to maintain or operate on the particular representation. The possibility that different representations might not actually be conflicting accounts, but might express different strategic approaches to reasoning was not considered.

Studies of problem solving that seemed to indicate use of different representations do not fit easily with unitary accounts of the FRM, these studies are reviewed in Chapters 2, 3 and 4. Roberts (1993) has eloquently articulated the problem that strategies have on attempting to pursue the FRM. He suggested that the existence of individual differences is the major challenge to such single representation accounts of the FRM. There are three possibilities for the impinging of strategies on FRM investigations. Either the representations exhibited by use of different strategies are interfering with the FRM's processing. Alternatively, different strategies will have the same underlying representations, but the underlying representations are overshadowed by the strategies. The final alternative is that the representations used by different strategies are not underwritten by any more fundamental mechanism – strategies are all we have. In any of these cases, deducing the FRM proves to be impossible: it cannot be observed directly, or it does not exist. However, studying the development and use of different strategies is the way to advance research in reasoning.

Strategies have been defined as "any procedure that is non-obligatory and goal-directed" (Siegler & Jenkins, 1989, p.11). Strategies in the sense discussed by Roberts, and adopted here, differ in terms of the representations that they invoke.

Representations are particular data structures. Newell provides a “representational law” for (internal) representations:

In an external world, entity (X) is transformed (T) into entity (Y). A representation of X-T-Y occurs in a medium within some system where an encoding from X to an entity in the medium (x) and an encoding of T into an internal transformation in the medium (t) produces an internal entity (y), which can be decoded to the external world to correspond to Y (1992, pp.426-7).

Understanding the representational system, then, requires an exposition of the entity (X), the transformation (T), and the operations on the entity (Y). Coupling a representational system with a procedure for using the representations defines a strategy.
Stenning and Oberlander (1995) present a complementary perspective on difficulties of postulating a unitary account of reasoning performance. They argue that the debate about the FRM has been conducted at the wrong level of description. Roberts proposes that the level of description should be in terms of the strategy - representation relationship. Stenning and Oberlander, however, argue that the point at which to discriminate different accounts is in terms of the representation - algorithm relationship. The FRM debate is misguided because it is impossible to discriminate the different accounts in terms of performance. This is because performance is in terms of input-output pairings and any number of representations could be used to implement this pairing. Stenning and Oberlander develop an Euler's Circles method that is indistinguishable from the mental models account in terms of the data used to previously support particular representational systems. Stenning and Yule (1997) further hammer home the point by developing a "natural deduction" method that is also indistinguishable from mental models. Each of these methods implements the same underlying algorithm, but the implementations differ. Stenning and Oberlander's challenge is to provide accounts of reasoning that explore the representation - algorithm relationship.

This thesis differs from Stenning and Oberlander's account in that different representations are seen to relate to different algorithms (even if these are not distinguishable by input-output pairings), but the different algorithms do have a strong family resemblance. Stenning and Oberlander suggest that use of different representations will inherit properties of the larger representational systems from which tokens are drawn, but that discriminating the influence of these systems requires more detailed data to individuate different accounts of reasoning. The differences between algorithms in the account proposed in this thesis are related to the properties of these larger systems.

The difference between algorithms and strategies, then, may be due to disciplinary boundaries: strategies and representations are within the domain of psychometric research, whereas representations and algorithms are discussed within computational accounts. Satisfying Newell's requirement on describing representations and coupling this with an account of procedure provides something very close to an algorithmic account. Exploring the strategy - representation - algorithm triangle forms the conceptual basis of this thesis. There are differences in levels of description here, but cognitive science benefits from corresponding concepts within different levels. Indeed, providing a full account of a system performing a task requires descriptions at several
levels of description. Marr distinguishes several levels for consideration:

At one extreme, the top level, is the abstract computational theory of the device, in which the performance of the device is characterized as a mapping from one kind of information to another ... In the center is the choice of representation for the input and output and the algorithm to be used to transform one into the other. And at the other extreme are the details of how the algorithm and representation are realized physically (1982, p.24).

Anderson (1989) finesses this levels account somewhat, by redefining the "computational" level of Marr's account with a "rational level". This frees up use of the "computational" level to refer to several different levels of description. Marr's use of "computational level" referred only to the goals of computation, and a richer conception of this level, including descriptions of the procedure utilised, would correspond better to studies of strategies, representations and algorithms. The rational level is intended to encompass descriptions in terms of goals and processes. Anderson also distinguishes Marr's representation and algorithm level into two separate levels: an algorithm level and an implementation level. The algorithm level is described as the "program symbol level" (p. 4, Anderson, 1989), where the task can be described in terms of symbol manipulation, or input-output pairings. The implementation level is the "register transfer level", which stands as an approximation to the biological level.

Neither Marr's nor Anderson's formulation of the different levels of description fully capture the properties of the strategy level, representation level, and algorithm levels of description required for interpreting reasoning tasks. The algorithm level correctly describes the input-output pairings of the processing underlying task performance, but these algorithms are implemented in particular systems of representation, which means something different to the grounding of symbols in a physical system. The embedding of the algorithm within a representational system is at a higher level of description than the algorithmic level. The rational level relates closely to conceptions of descriptions of strategy differences. Newell's formulation of the principle of rationality is that "if an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action" (p.102, 1982). The knowledge level places constraints on actions, it does not entirely determine them.

The psychometric approach has had a slightly different agenda. The history and methodology of psychometrics is now discussed, before the way psychometric and computational accounts are shown to be mutually compatible and beneficial approaches.
1.2 The psychometric approach

Many of the strategies discussed in this thesis utilise either graphical or verbal representations to solve problems. One approach to study use of these different representations is to relate the selection or effectiveness of their employment with measures of “cognitive ability” or “cognitive style” for using these different representations.

Cognitive ability can be defined as “a person’s performance on some task that has a substantial information-processing component” (p.93, Cooper, 1990). Messick defines them similarly: “Dimensions of intellective ability refer to the content, component processes, and level of cognition - to the questions of what? and how much? What kind of information is being processed by what operation in what form and how well?” (Messick, 1984, pp.62-63). The use of ability measures is one psychometric approach to studying complex task performance. Spearman (1904) gave subjects a set of ability tests, and performed factor analysis on test performance. He determined a single factor $g$: a measure of “general” ability, or “intelligence”. Thurstone (1957) gave a similar set of tests to subjects, but identified 12 separate factors. These “primary mental abilities” included spatial ability, verbal relations ability, and (not very usefully, for the current studies) deduction ability. Individuals vary in the extent to which they have a particular ability. Once identified, the psychometric research program suggested that such basic abilities can be found to be related to any cognitive task, insofar as the task requires the processes of the various sub-factors of $g$. The development of items on such tests, though, is rather ad hoc, and it is often not clear what the tests are assessing or whether subjects are solving the tasks the same way. These problems are further discussed with regard to spatial ability measures in Chapter 2.

An alternative psychometric approach is to assess cognitive styles. These are measures of the propensity or preference an individual has for processing information in a particular manner. Unlike abilities, styles are morally neutral. Being at one end of the style scale may reflect better performance on certain tasks but worse performance on other tasks. Different styles may be more or less appropriate as the task varies.

Relating complex tasks to psychometric measures, then, applies descriptions of performance at different levels of description. Strategy choice can be related to style, insofar as the different strategies used are predicted by some style measure. Strategy choice can also be related to the abilities underlying the use of a particular strategy: if ability measures relate to preferential se-
lection of strategies that use the processes assessed by the ability measure then this suggests that there is some causal connection between them, so abilities are intended to expose the processes underlying representation use.

There is a distinct difficulty with using the psychometric approach to determine the processes involved in reasoning performance. Finding a test that probes a particular primary ability and no other has proved difficult. Indeed, approaches that have attempted to produce measures of purity have then had the unfortunate consequence of just not being very interesting, as they do not relate to performance on any complex tasks! This difficulty means that psychometric studies tend to use many subjects and many tests in order to reduce the noise in the measures and to factor out the “primary” abilities. Specifying the task requirements in greater detail, and expanding on the parameters of performance is one way to address this problem. Then the “noise” can be made the topic of study – in any case, this may well be the point of interest in performance. Similarly for styles, without expressing the connection between style and the properties of the representations used and the procedure operating on the representations will produce an incomplete picture of performance. Without this information-processing supplement provided by computational approaches, the best that can be achieved is correlation without explanation.

1.3 Integrating individual differences

The essence of cognitive science is to explore and extol the virtues of the overlap between disciplines. So, how will cognitive science benefit from psychometric research? As a point of departure, relating psychometric research to computational accounts will provide some justification for observations of individual differences in tasks typically within the domain of computational theories. Noting their existence is a preliminary to providing justification for why computational accounts ought to take them into consideration.

How will psychometric research benefit from the computational approach? Johnson-Laird wrote:

Whatever the general merit of investigating ‘individual differences’ by way of mental tests, their use is unfortunately of little value in the study of reasoning. The data they yield are ... too gross to elucidate differences in mental processes from one individual to another” (p.117, 1983).

To address this criticism, psychometrics have to be more than mere snapshot measures of performance. They have to contribute towards accounts of process. Many psychometric measures are
intended to assess response to tasks that require information-processing in a variety of modalities and different operations on representations within those modalities. Coming to terms with the computational properties of the representations will assist in providing explanatory accounts of the connections between different psychometric measures over and above descriptive accounts. The measures themselves can be subjected to computational investigation, and this enriches understanding of the reason behind the connection between measures and reasoning tasks.

Johnson-Laird (1983) expresses a problem with using individual differences which suggests one reason why cognitive science has been resistant to contemplating variation in response. In criticizing an attempt by Guyote and Sternberg (1981) to relate syllogism solution to measures of spatial ability and verbal ability, he refers to their approach as attempting to provide "a model of actuarial data" (p.87) rather than giving an account of individual minds. The mathematical model provides us with an account of a "group mind" (p.86). A better approach is to provide a computational account of the processes involved, with basic parameters that can be adjusted according to individual differences, such as resource availability, working memory span, or rate of degradation of information. This approach is exemplified by Johnson-Laird's mental models account, or by ACT-R (Anderson, 1993) or SOAR (Newell, 1990) cognitive models.

But the simplifying assumption taken in cognitive science that there are basic parameters that can be adjusted to account for difference in performance is as primitive as the approach of Guyote and Sternberg. These accounts appear "gross" when qualitative rather than quantitative differences between individuals are observed in task performance. Parameter-shifting alone cannot account for the sizeable differences in performance that accompany use of different representations to solve reasoning problems, for example. When styles are at issue rather than abilities, this presents a challenge to unified theories of problem solving. This is not to say that low-level parameters do not influence the selection and use of high-level strategies, and a central aim of this thesis is to indicate that several supposedly qualitative differences, such as representational modality, can be interpreted along a continuum when described computationally. But absent from many computational accounts of performance are the necessary degrees of freedom for modelling these style differences at all levels of description.

Doing away with the simplifying assumption of homogeneity in response is akin to making the variance of the population, rather than the mean, the topic of study. This does not necessarily mean a more complex theory, but rather entails a richer understanding of the range of responses
and gives an indication of the size of the solution space for problems. Newell writes: “The prime question about computational systems is what functions they can produce” (my emphasis, Newell, 1992, p.427). Studying a single approach to problem solving, for example, will provide an account of what function a system is producing. To address the potentiality of the system, one needs to study a variety of strategies. Determining the range of strategies will, at least, begin to address the issue of defining the range of possible solutions, rather than just plotting individual points.

Chater and Oaksford’s (1999) rational analysis approach has indicated the importance of variation in response as revealing the local minima surrounding the optimal solution. Their account of the four-card task illustrates the power of considering variations in response in order to finesse understanding of the task and the range of possible performance. In the four-card task subjects are presented with four cards, each of which has an A, a K, a 4 or a 7 on the side facing up. The subject is told to turn the fewest cards they have to to test whether the rule: “If there’s a vowel on one side there’s an even number on the other” is true or false. Most subjects respond by turning the A and the 4, or just the A card. Very few subjects (about 4%) turn the A and the 7 cards, which is the normative answer. Oaksford and Chater (1994) found that an optimal data selection analysis predicted an order to turning the cards. If a subject will turn just one card, then they will turn the A, if they turn two, it will be the A and the 4, and so on. This approach accounts for the majority of responses, but it does not explain how and why some subjects respond with the A and the 7 cards. These subjects have to be solving the task using a different strategy, which would require an alternative “rationality”.

This is the approach of Stanovich (1999) who argues for a dual-process theory of reasoning, distinguishing individuals in terms of the extent to which they decontextualise information. In the four-card task, for example, the optimal data selection analysis will only apply if information is embedded in a context, a context where co-occurrence (in the four-card task, the co-occurrence of vowels and even numbers) is rare. If information is decontextualised, then different principles underlying data selection are at play. This is a promising line of investigation, conducted by Stanovich and colleagues within the psychometric domain. A computational account of decontextualisation will enable the alignment of their stylistic account with such rational analysis approaches.

Aptitude treatment interactions (ATIs) describe the relationship between a subject’s profile as
measured by psychometrics and the extent to which they learn from different teaching interventions. Such phenomena, though rare, present a powerful challenge to unified computational accounts of performance, as they cannot then be dismissed as noise. They also present an important area of overlap between psychometric and computational studies of cognition. A fuller understanding of the nature of ATIs has implications for teaching of problem solving as well as the design of educational programmes in general. These ATIs are the principal point of synthesis for this thesis.

1.4 Outline of the thesis

Chapter 2 is a review of studies of reasoning tasks that exemplify a range of strategies used to solve the problems. These strategies are reinterpreted in a framework that considers the computational properties of the representations used by each strategy, rather than describing the strategies just in terms of their apparent use of different representations.

Chapter 3 sets the psychometric context, by considering several measures that are potentially useful for discriminating "ability" and "style" dimensions relevant to representational and strategy differences in reasoning. Requirements for suitable tasks for relating psychometric and computational approaches are discussed, and illustrated in a preliminary study of strategy variation and change in Hyperproof.

Chapter 4 takes up strategy and representation variation in syllogism solution. Syllogisms have been closely studied, and computational theories of their solution abound. As the introduction indicated, this is also an area of bountiful confusion, with different levels of description being confounded. Accounts that assess differences in strategy, representation and algorithms are consolidated.

Chapter 5 summarises the aims and hypotheses of the empirical studies, studies that attempt to align psychometric with computational accounts of individual differences in reasoning.

Chapter 6 then presents a study on teaching syllogisms using different representations, providing empirical support to the theoretical distinctions offered in Chapter 4. This study explicitly connects psychometric and computational accounts of reasoning.

Chapter 7 presents a replication of the teaching study in a different domain, returning to Hyperproof. This study relates responses to taught methods to spontaneous strategy development. The
strategies devised by subjects are related to the psychometric and computational interpretations of representational differences used by the different strategies. This study also addresses issues of transfer of skills between domains, providing a computational perspective on observed ATIs.

Finally, Chapter 8 summarises the main achievements of the thesis, and addresses some potential criticisms of the account. The empirical studies presented are embedded in a larger program that uses individual differences to reveal the range and the fundamental nature of cognition.
Chapter 2

Strategies in complex problem solving

2.1 Strategies in reasoning

Responses to reasoning tasks are replete with strategic variation, but is this variation ad hoc, changing from task to task and from subject to subject, or is there some commonality and generalisability in the strategies employed? There are several observations of the general nature of strategy change, which support the pursuit of an account that generalises over different reasoning domains.

The aim of this chapter is to describe the range of strategies on a set of established reasoning tasks, and the manner in which strategies develop in a principled way. The principle underlying such a description is in terms of the level of abstraction in the representations used by the strategy. This description is grounded in computational theories of the difference between graphical and verbal forms of information in terms of the expressivity of the system from which the representations are drawn, indicating a first pass at the strategy - representation - algorithm distinction. Though the difference in expressivity may be slight for the alternative strategies presented here, in Chapter 4 the same distinction is illustrated for syllogisms where the differences in expressivity is greater.

Before moving on to describe this computational framework, the common characteristics of strategies in reasoning domains are discussed, and several studies which suggest common mech-
anism are reported.

Roberts (1999) attempted to classify and generalise different strategies that are used in a range of problem solving tasks. Roberts identified three types of strategy that subjects use on a range of reasoning tasks:

- **spatial strategies**: "information is represented spatially, such that the configural information in the representation corresponds to the state of the affairs in the world" (p.4);
- **verbal strategies**: "information is represented in the form of verbal or abstract propositions, and that various content/context-free syntactic rules enable new conclusions to be drawn from the represented information" (p.5);
- **task-specific short-cut strategies**: result when subjects “notice certain regularities or redundancies during the problem solving process” (p.5). Quinton and Fellows (1975) termed such strategies representation-free strategies.

Roberts (in press) observes a general change across a number of reasoning domains from spatial to verbal strategies – several domains where such changes occur will be discussed below. One aim of this chapter is to recast such changes in terms of the computational properties of different representations of information, using the specificity/expressivity distinction (Stenning & Oberlander, 1995). This enables a clarification of the difference between spatial and verbal information presentations, which avoids recourse to phenomenology (Pylyshyn, 1973; Anderson, 1978) and also marks the short-cut strategies as an extreme on the spatial-verbal dimension used to make distinctions between strategies. After detailing the specificity/expressivity theory and discussing how it relates to spatial, verbal and short-cut strategies, three problem solving domains that have indicated strategic variation in solution strategies will be reinterpreted in this framework. Previously, the specificity/expressivity framework has only been applied to the use of static representations, here it is extended to apply to dynamic features of strategy development and change.

Several descriptions of the nature of the change in process determined by strategic change have been registered.

---

1Page numbers are taken from a pre-publication manuscript.
Changing strategy in the linear syllogism task results in different use of memory resources: "The subject supposedly moves from a reliance on 'symbol' memory to a more positionally governed technique where merely to read through statements in a rule-governed sequence is to solve the problem" (Wood, 1978, p.333). Positing different uses of memory resources as a feature of different strategies is a widely applied account. Baddeley's (1990) model of working memory includes separate storage modules for spatial and verbal information. However, whether the different memory traces are best described in terms of the representations (as is currently the case) or processes is a point returned to in the next chapters. Wood's (1978) account of different uses of memory may be that the list method uses memory storage whereas the search method does not use any memory trace, though it is hard to believe that no memory storage is required by the latter method, in terms of keeping track of which information has already been searched, for instance.

A second description of the general nature of strategy change is made by Wason and Johnson-Laird (1972). They describe the strategy change for students that have protracted experience with tasks as a development from representational to non-representational thinking. They write:

... the inexperienced subject represents the premises in a unified form (with or without imagery) because this is likely to be the normal practical mode of dealing with the relational information. But by dint of sheer repetition this approach is likely to give way to a purer and more formal strategy geared to the specific constraints of the problem (p.122).

Practice in a task leads to a more "mechanised" approach "which minimizes effort and which is appropriate to the particular constraints of the material" (p.126).

Relatedly, Wood (1978, p.336) describes the development of strategies in terms of "economy". For example, in changing from search to scan strategies in linear syllogisms (see below), if the subject searches for "A is taller than B" and this proves unsuccessful, then any searches after this point are redundant as a consequence of the task constraints is that necessarily "B is taller than A". The subject's realisation of this means the method can be pruned.

All these accounts utilise terms like efficiency, economy, and specificity to describe the development of strategies. These observations prepare the ground for a theory of the information used by different strategies that appeals to how the different types of information provide differing degrees of specificity and efficiency in processing. The next section discusses representations in terms of their informational properties.
2.2 Spatial and verbal: specificity and expressivity

Larkin and Simon (1987) consider the computational differences between diagrammatic and sentential representations for solving problems when these representations are informationally equivalent. Their analysis concerns external representations, but the processes that operate on these representations are internalised. They consider the critical properties of sentential representations to be that the data structure is such that elements appear in a single sequence, whereas for diagrammatic representations the data structure is organised such that information is indexed by two-dimensional location (p.68). These differences in representation cause differences in the facilitation of solution for a range of problems in a variety of subject domains, such as mathematics, or economics. The informational differences induced by the data structures fall into three categories: Search is a process where sets of elements are located that satisfy the conditions of one or more productions (p.69). For sentential representations, the search must advance through the list. For diagrammatic representations, the search may be localised, where elements required for the inference are grouped together. The second category is differences in recognition which are due to the explicitness of the diagram, where inferences are immediately apparent through the process of drawing the diagram. For example, being told that two lines intersect additionally provides data about four angles when a diagram is produced, this is information that remains implicit in a sentential expression. This property is due to the specificity of diagrammatic representations, where decisions about the inter-relationships between elements of the problem have to be made and are demonstrated in terms of the relative location of elements in the representation. The third category – inference differences – is seen to be less strong, indeed, Larkin and Simon claim that inference is largely independent of representation if the information content of the two sets of inference rules is equivalent (p.71).

Larkin and Simon pick up on the property of specificity of spatial representations. However, their discussion concerns the specificity of individual tokens of information in graphical/spatial representations. This account is difficult to apply in the case of strategy change, because the researcher needs to know exactly which tokens are used in each strategy. A more far-reaching account would address informational properties of the system from which the tokens are drawn, and this is the approach taken by Stenning and Oberlander (1995). They observe:

graphical representations such as diagrams limit abstraction and thereby aid processibility. We term this property of graphical systems of representation specificity: the
demand by a system of representation that information in some class be specified in any interpretable representation. (p.91).

Whereas Larkin and Simon contrast computational and informational properties, Stenning and Oberlander contrast the logic of a task with its implementation. The logic of a task is the underlying abstract algorithm, and the implementation is the specific representation of the algorithm, which is committed to a certain media of presentation.

Stenning and Oberlander (1995) use Euler’s Circles to illustrate this contrast and exhibit the computational properties of graphical representations. Euler’s Circles were useful for Euler’s pupils, both past and present, because of their combination of requiring certain information to be specified, together with only allowing limited abstraction. Systems where abstraction is too limited are not so powerful, and systems where abstraction is unrestricted prove difficult to operate and apply to a given problem. The specificity of a system exists at the expense of its expressivity, and finding the appropriate level of specificity is the ideal for solving tasks with the greatest ease. This is a motif that will be returned to continually in discussing strategies used in reasoning tasks.

Stenning and Yule (1997) devised a verbal system for solving syllogisms, which was designed to be equally expressive as the Euler’s Circles system developed by Stenning and Oberlander (1995). However, the general languages from which these two systems draw their tokens are different in their expressivity. The system from which the verbal system is drawn can express more relations than that from which the Euler’s Circles method takes its tokens (see Chapter 4).

The expressivity/specificity distinction immediately suggests application to the literature on strategic variation in deductive tasks. Most apparent is that it deals with differences between spatial and verbal information, and so redescriptions of strategies that invoke such representations thus benefit from this informational interpretation. The change from general to specific strategies mirrors the change from creating “unified” representations to more economical representations (Wason & Johnson-Laird, 1972), which means that the representations are increasingly abstract, and drawn from increasingly expressive systems. In complex problem solving tasks, subjects often seem to overshoot in the specificity of their representation, meaning that problem solving is less efficient before the appropriate level of expressivity for the task has been found. This description of change in terms of expressivity has the advantage of being modality independent, which explains why in some tasks observed change is from spatial strategies to verbal strategies,
and in others the change is in the opposite direction. In computational terms, specifying all and only the information required for solving the task means greater tractability. The following discussion of the variety of strategies used for reasoning tasks indicates that the subject’s ability or propensity to hone down their system to this minimum is a matter of individual differences, as is the student’s preference for using representations that invoke various levels of abstraction. Use of different levels of abstraction in representation relates to different algorithms for solving the problem, but these representational preferences can equally be expressed as strategic variations.

In the next section, three deductive reasoning tasks are described in detail, indicating general features of strategy variation and development in terms of changing levels of abstraction in the representations used.

2.3 Expressivity/specificity in reasoning

2.3.1 Linear Syllogisms

Linear syllogisms are sets of statements about particular terms that define the relationship between the terms. The relationship is asymmetrical, and usually about the relative size or position of terms. The type of linear syllogism that Wood (1969; 1978) investigated was of the following form, though the number of statements can vary (for example, Sternberg (1980) and Sternberg and Weil (1980), discussed below, studied problems with only two statements):

John is taller than Paul
Dave is taller than Paul
Tom is taller than Dave
Ian is taller than Tom
Dave is taller than John

The subject is asked to say who is taller from Dave and Ian. The presentation of the problem requires certain inferences in the material to be made explicit (Wason & Johnson-Laird, 1972): in the above case in order to solve the problem the subject needs to apply knowledge about the asymmetry and transitivity of the relationship “taller”. One strategy for solving this task, described by Huttenlocher (1968), is to construct a “ladder” where tall items occur at the top of the list and shorter items occur at the bottom. The representation for the above problem is shown here:
The spatial arrangement is not essential as a feature of this strategy, only that the relations between all of the individuals are represented in some way. Then the answer is "read off". This strategy expresses all the relations between individuals in the problem, where all inferences are made explicit in constructing the representation. The spatial strategy produces the "unified" form, or the representation that makes concrete all the information in the problem. Therefore, this can be seen as the strategy using the most concrete form of representation.

Once the subject has had lengthy experience with the task, Wood (1969) found that many subjects report using a different strategy, where a selective search is performed. The first term in the question is taken and a search performed down the left column of the statements. If a match occurs, then the subject checks to see if this connects it to the second name in the question. If so, then the answer is available, if not then they search for occurrences of this new item to see if this links it to the second term in the question. If there is no link, then subjects begin a search with the second name. With more experience of the task, some subjects come to realise that if there is no link from the first name, then the second must be the taller, and this reduces their solution time by 25% (Wood, 1978, p.332). This strategy does not make explicit all the relations between terms in the syllogism, only the relations between relevant terms are expressed. For the above problem, the representation used by this strategy is as follows (again, the spatial arrangement is not important, the feature in question is the extent to which the relations between names are expressed):

Ian
Tom
Dave

For this problem, the relative size of the other two items is not specified, and so the representation used by this strategy is more abstract.
A third strategy has been observed by Wood (1969). Some subjects adapt their search so that they scan the left-hand side of the statements in the list. For some problems, this determines if one or other person cannot be taller as if they don't appear on the left they can't be taller than anyone. If the question for the above problem was "Who is taller, Peter or Tom", then the subject could scan the left of the list for the name Peter, find his name does not appear there, and conclude that "Tom is taller". In this case, the representation used would only be explicit about the relative size of one item:

...all others ...

Peter

Thus, three strategies have been observed for this type of problem: list, search, and scan strategies. Wood (1978) reports that when a subject uses more than one strategy, there is always a "privilege of occurrence", so that the list method comes before the search method which precedes the scan method. Later strategies were found to be more efficient. Wood (1978) found that subjects who are initially very good at solving linear syllogisms using list strategies are those that develop "short-cut" strategies sooner. Relatedly, in a teaching study, Quinton and Fellows (1975) found that subjects performed most quickly and accurately if they were taught to use the scan method, and least quickly and accurately when taught to use the list method.

The most efficient strategy is one that uses the most appropriate level of abstraction for the task in hand. The list strategy makes more information explicit than is necessary, and this means that much effort is expended in forging the inferences to work out the relations between all the items in the problem, and then the storage of all the terms is additionally effortful. A more efficient method will only specify the relations between terms that are required in order to reach the solution: so either the two terms in the question and those items that intervene, as in the search method, or just one term, as in the scan method. Hence, there is a distinct connection between the specificity of the representation (in terms of how task-specific it is), its efficiency, and its "privilege of occurrence".

Wood's studies report that the more able subjects tended to discover the more efficient strategies sooner, they seemed to be better attuned to the constraints of the task that make such strategic discovery possible. This suggests that there is either an ability or a style dimension relevant to the development of strategies that vary in terms of their specificity.
An alternative, complementary investigation into linear syllogistic problem solving was conducted by Sternberg (1980) and Sternberg and Weil (1980). They investigated linear syllogisms of the form:

John is taller than Bill
Pete is not taller than Bill
Who is tallest?

The statements can vary in terms of whether they state a relation of "taller than", "shorter than" or of "not taller than" or "not shorter than". Their aim was not so much to examine strategy change as students solved the problem, but to investigate variation in initial strategy use and the effects of teaching strategies for solving these problems.

Though there is less variation possible in terms of specifying or leaving abstract the relations between terms, Sternberg (1980) has provided componential models for four different strategies that were used to solve these problems. These strategies vary in terms of the abstraction present in the representations.

The "spatial" method arranges the terms of the syllogism spatially "in an imaginal, linear array that is an analogue to a physical, linear array" (Sternberg & Weil, 1980, p.236). Each premiss is used to construct a separate array, with the properties flipped in the array if a negation is present in the premiss. Then the two arrays are integrated, the question read, and the array examined for the answer. This is very similar to Wood's list strategy, and, as with Wood's analysis, is also related to Huttenlocher's (1968) spatial strategy model of subjects' performance. So, for the above example, the following representation is constructed:

John
Bill
Pete

The "linguistic" method stores the terms of the syllogism using functional relations in terms of the deep structure of the premiss statements, thus properties have one of the predicates of tall, tall+, short, or short+. If negation occurs in a premiss, then a linguistic transformation changes the propositional statements into one of the four forms above. Information is left unintegrated, but the middle term of the problem is searched for, then the question is read and a final search
of the propositional terms is made. So, the representation in use only specifies the relative size of the middle term and one of the other individuals:

John
Bill

This is akin to Wood's search strategy in that it extracts only the information that is required for the search, avoiding the additional effort required for integrating the information. The two statement version of the task means that search is for the middle term whereas in the multiple statement task search must start with one of the target terms and investigate the intervening terms before finding the other target term. Hence, it can be seen as a special case of the search strategy, one which is only feasible due to the additional constraints of the two statement task.

Sternberg's third model – the "linguistic-spatial mixed" method – decodes the premisses into linguistic propositions, but then reorders these propositions into a spatial array. As with the linguistic model, the middle term must be identified from the linguistic encodings in order to construct the array.

Finally, in the "algorithmic" method the question is read first then the question answered in terms of a linguistic encoding of the first premiss. Then, the second premiss is scanned ignoring the relationships between the terms. If the answer to the first statement is not contained in the second statement then the first answer is taken. If the answer to the first statement is contained in the second premiss then the answer is the other possible answer choice in the second premiss.

This strategy makes more use of the constraints of the task than did the linguistic strategy, so it can be seen as the most task-specific of the strategies described by Sternberg (1980), or as a shortcut strategy in Roberts' (in press) framework. In the above example, only the information of the relative position of Bill will be specified. This strategy is comparable to Wood's scan strategy, in that it only attempts to garner information about one of the items.

In Sternberg and Weil (1980) support was gathered for the implementation of these strategies in human performance. In addition, an aptitude treatment interaction was assessed, where students were pre-tested for verbal and spatial ability and then taught to use either the spatial or the algorithmic method. The pre-test verbal measures were a word grouping task where subjects had to indicate which word does not fit with the other four, and Form S of the DAT Verbal Reasoning Test which is a verbal analogy test. The first spatial test was Card Rotation from the French Kit of
Reference Tests for Cognitive Factors, where two-dimensional shapes at different angles have to be compared as mirror-images or identical. The second spatial test is Form S of the DAT Spatial Relations Test, which requires a shape to be folded from a two-dimensional outline and compared to three-dimensional shapes. After solving 40 linear syllogisms, subjects took the pre-tests and were either instructed in the spatial method, the algorithmic method, or had no instruction. They then solved a further set of syllogisms.

The results indicated that those taught to use the algorithmic strategy were faster at solving the post-instructional problems (Newman Keuls test, $F(2, 141) = 25.91$, $p<0.001$), but the spatially instructed students did not differ from uninstructed subjects. This supports the results from Wood's studies showing that strategies specifying less information are more efficient for this task.

Fitting each teaching group against the predictions from each of the four theoretically motivated models indicated that both the untrained and the spatially taught group used a mixed model, but the algorithmic taught group did not fit the pattern for any of the models. Pre-test measures were combined into a spatial factor and a verbal factor. Significant negative correlations between these factors and solution latencies for each teaching group were significant for all groups for every factor.

Re-sorting groups according to their model fit on the post-instructional problems indicated that many students did not use the method in which they had been instructed. When assigned to groups according to the strategy determined by a best fit of the latency data for each individual, a significant negative correlation between solution latencies for the mixed method group and both spatial and verbal factors was found ($r = -0.27, p<0.01$, $r = -0.45, p<0.001$). The linguistic model group latencies negatively correlated significantly with the verbal factor ($r = -0.76, p<0.001$) but not with the spatial factor. The spatial model negatively correlated with the spatial factor ($r = -0.61, p<0.01$) but not with the verbal factor. The algorithmic model correlated negatively with the verbal factor ($r = -0.32, p<0.05$), but not with the spatial factor. Hence, an aptitude treatment interaction was found when subjects were classified according to the strategy they seemed to be using. Spatial and verbal ability measures are found to relate to effective use of strategies that use spatial and verbal representations, respectively.

Sternberg and Weil (1980) show that a wide variety of strategies, varying in terms of their expressivity are used, and can be taught, though with limited success as subjects often used a strategy
other than the one they were taught to employ. Unfortunately, the data of the studies does not indicate the relationship between strategy selection and the pre-test measures. It does address the issue of effective use of a strategy once it has been selected, and indicates that subjects with high spatial ability are better able to use a method that specifies and resolves more information. Subjects with high verbal ability are better at using all the strategies.

Wood’s approach explores strategy change and shows that there is a general move from strategies that draw on specific systems to strategies that draw on more expressive systems, though difference in expressivity is only slight for this task. This difference in expressivity can equally be expressed in terms of the level of abstraction present in the representations. Wood’s more complicated version of the task seems to promote a wider variety of strategy, and the greater processing load in the spatial task seems to provoke strategic development in order to reduce the load. The relationship between strategy use for solving linear syllogisms and psychometrics has not been extensively studied and the Sternberg and Weil (1980) results indicate correlations with performance without explaining how ability helps for each method, or how strategies are selected. Another domain where the relationship between psychometrics and strategy use has been investigated is the sentence-picture verification task, where spatial ability has been used to reflect strategy selection.

### 2.3.2 Sentence-picture verification

The sentence-picture verification task was originally designed to assess the ways in which negation was processed in sentences (Wason, 1972; Wason & Jones, 1968). It involves the subject viewing a sentence followed by a picture, and assessing the truth value of the sentence with respect to the picture. Clark and Chase (1972) and Carpenter and Just (1975) described a method of solution in which the picture is translated into a sentential form and then compared with the sentence. This was termed a “verbal” strategy. Thus, a sentence containing negations would take longer to process than a simpler sentence. Response time data supported this description. However, for Carpenter and Just’s study, only the first few trials were analysed, as pilot studies indicated a systematic change in response times after being exposed to many problems. As Roberts, Wood and Gilmore (1994) comment: “This leads one to wonder exactly what the status should be of an all-embracing model of performance that only applies to the first 36 trials of a 252 trial task” (p.416).
Macleod, Hunt and Mathews (1978) found that the performance of about a quarter of their subjects did not fit the expected pattern of processing time associated with the verbal strategy: these subjects did not take longer over more complex sentences containing more negations as the verbal strategy would predict. Macleod et al. proposed that these subjects were using an alternative "spatial" strategy where the sentence is translated into a spatial array, which is then compared with the picture. Use of these different strategies (verbal or spatial) can be detected by different lengths of processing of the stages of the task. A subject using the spatial strategy will take longer over processing the sentence stage, and less time over processing the picture stage than a subject using the verbal strategy.

Marquer and Pereira (1990) found that introspective accounts from subjects while they were performing the sentence-picture verification task did not match the response time data corresponding to the strategy the subject said they were using. The majority of subjects whose data fit the spatial strategy reported using verbal strategies. Roberts, Wood and Gilmore (1994) suggested that a verbal recoding strategy would produce the same response time data as the spatial strategy. This strategy involves recoding negative statements as affirmatives and then this sentence is compared to the picture. Like the spatial strategy, this produces response times that are invariant with regard to whether negation is present or absent. Roberts et al. term this type of strategy a "flat strategy", a characterisation independent from representation type, instead invoking level of abstraction in order to describe the different families of strategy.

How do the verbal and the flat strategies differ in terms of the level of abstraction used in the representations? For the verbal strategy, the information is more abstract in that the encoding is also compatible with situations where the shapes are alongside each other – it is the maintenance of negation in the encoding that gives this, and this means that the representation system is more expressive. In the verbal strategy, there are more situations consistent with the expression used. In the spatial strategy, the shapes are specified in terms of one always being above the other, so the transformation from, say, "The cross is not below the asterisk" to "The cross is above the asterisk" is performed in the spatial strategy, but not in the verbal strategy. The spatial strategy specifies a particular situation, and so the representational system is less expressive.

Two features of strategy variation raised with regard to the linear syllogism studies are now investigated. The first pursues the issue of strategy development, and the second looks at the relationship between psychometrics and strategy use.
For the sentence-picture verification task, the change in representational form seems to be from more abstract to more concrete. Carpenter and Just (1975) mention different performance in later trials on the task, which were incompatible with a verbal encoding model. This suggests some change in the expressivity of the representational system occurred after exposure to the task, and a change away from the verbal strategy that uses more abstract representations. This is in the reverse direction to the change observed in the linear syllogisms study, where the privilege of occurrence changes from more concrete to more abstract representation use. But is the flat strategy more efficient than the verbal strategy? If so, then this means strategy change in both the linear syllogism and the sentence-picture verification task is towards more efficient or appropriate levels of abstraction. Macleod et al.'s (1978) study does not indicate that the flat strategy (in their study, described as a spatial strategy) is more efficient than the verbal strategy, if response time is taken as a measure of efficiency. Verbal strategy subjects took an average of 2862ms to respond, whereas the spatial strategy subjects took 3230ms on average. In Roberts, Wood and Gilmore's (1994) study, no difference in response time is observed, either. For the flat strategy, mean RT = 2.51s, and for the verbal strategy, mean RT = 2.48s.

However, two types of problem can be distinguished: those where the sentence is affirmative, and those where the sentence is negative. It is predicted that the former are easier for the verbal strategy, but the latter would be more difficult for the verbal strategy, as the representation used is too abstract to be optimal. Response times for problem type by strategy group from the subjects in Roberts et al.'s (1994) study are reproduced in Figure 2.1.

These results suggest, as anticipated, that the verbal strategy is better for the affirmative problems, but worse for the negative problems. An ANOVA of response time with strategy and problem type as factors was performed. A main effect for problem type was found (F(1, 39) = 112.85, p<0.0001) with affirmative problems more quickly solved. There was no main effect for strategy. As predicted, there was a significant interaction of strategy with problem type (F(1, 39) = 7.72, p<0.01), confirming the suggestion that the flat strategy is most appropriate for the more difficult negative problems.

The second issue was the relationship between strategy use and psychometric measures. Macleod et al. (1978) found that subjects using the flat strategy responded more quickly if they

---

2Grateful thanks to Max Roberts for performing this analysis. The design is imbalanced, so results must be treated with caution.
had high spatial ability than subjects using this strategy with lower spatial ability. This reflects the Sternberg and Weil (1980) finding that high spatial ability related to effective use of the spatial strategy for linear syllogisms. In addition, spatial ability was found to relate to strategy selection as well as use. The mean spatial score for subjects using the flat strategy was higher than that for the subjects using the verbal strategy, but verbal ability was equal for each group³. This result, coupled with the response time data from Roberts et al. (1994) suggests that those with high spatial ability choose a strategy that is most effective for the harder problems. Their choice minimises the effort required for harder problems, at the expense of making easier problems slightly more effortful.

The compass directions task also demonstrates different representation use in problem solving, and studies have connected psychometrics to strategic variation and change for this reasoning domain.

³Mathews, Hunt and Macleod (1980) failed to reproduce this aptitude-selection relationship, so the results must be treated only as suggestive.
2.3.3 Compass Directions Task

Roberts, Gilmore, and Wood (1997) studied individual differences in response to the compass directions task (Wood, 1978). This task presents a series of directions simultaneously, and subjects have to say what point on the compass they end up at with respect to the starting point. An example is:

One step East
One step South
One step East
One step South
One step West
One step West
One step North
One step West

The answer to this problem is South-West (this being the nearest point to the finishing position on a 16 point compass).

One method of solving this problem is to draw or imagine moving each step, then calculate the answer from the resulting path. This is termed a "spatial" strategy, according to Roberts’ classification system. The path for the above example is illustrated on the left of Figure 2.2. It requires the starting point to be maintained, as well as every intervening point on the path, and
can thus be interpreted as a concrete representation. Alternatively, the starting point and position with respect to the starting point must be maintained.

An alternative method is to use a cancellation strategy, where moves East/West and North/South cancel each other out in the list until there are just the non-cancelled moves left. The cancellation strategy applied to the above problem is shown on the right of Figure 2.2. Counting the remaining moves after cancelling gives the answer: South-West. The cancellation strategy is a short-cut strategy: it is not generally applicable (consider a variant on the task where movements can be in any direction for varying numbers of steps). The cancellation strategy avoids the requirement that the initial or intermediate positions be maintained by making explicit the inverse operations of moving east/west or north/south that are implicit in the path of the spatial strategy. As the starting point and the intermediate steps are not represented this method can be seen as using more abstraction. As with the flat strategy in the sentence-picture verification task the amount of information that has to be specified in the representation is reduced.

Subjects reported using both the spatial and the cancellation strategy, and there was a general trend towards using the cancellation strategy on later trials (see also Newton & Roberts, in press). This suggests that the later developed strategy is one that employs operations to minimise the amount of information that has to be specified. This is closely akin to the strategy change that occurs in the linear syllogisms.

Strategy use was related to spatial ability in these studies. To measure spatial ability, subjects were given the Saville-Holdsworth Advanced Test Battery Spatial Reasoning Test (Saville & Holdsworth Ltd, 1979). Items on this test require subjects to imagine folding 2-dimensional figures to make patterned cubes in order to decide to which of several possibilities they match. Subjects with high spatial ability did not use the spatial strategy, but tended to use the cancellation strategy instead. This inverted aptitude-strategy relationship suggests that high spatial ability does not mean that spatial strategies are employed. When coupled with the sentence-picture verification task data, however, there is a suggestion that high spatial ability subjects use more abstract representations that in this case reduce the amount of information that has to be maintained by the subject. As with the sentence-picture verification task, high spatial ability subjects seem to use strategies that minimise effort, at least for harder problems.

Roberts et al. (1997) extended the initial study of the relationship between strategy selection and spatial ability by presenting an adapted version of the compass directions task. This adaptation
requires the subject to assess the final position of two individuals moving according to given directions. Moves were presented one at a time on a computer screen. Using the cancellation strategy on this method is particularly difficult as applying the cancellation rules requires storing the moves (this is not the case when moves are presented in parallel) and some moves must also be reversed before they can be cancelled. Roberts et al. (1997) refer to the cancellation strategy being more efficient in terms of working memory for the simple compass directions task, but less efficient for the two person compass directions task. They suggest this is due to differing loads on working memory resources but the precise details of the different loads are not specified. The harder task seems to mean that the extra effort of applying cancellation is greater than that required to construct a representation that includes the starting point and/or intervening steps.

High spatial subjects were found to be more flexible than low spatial subjects: they were more likely to change to the spatial strategy on this second task. When the task changes, high spatial ability subjects are sensitive to the appropriateness of the strategy for the task and switch accordingly. The development of the cancellation strategy for the standard compass directions task only occurs when there is some memory load, however. When subjects were given paper and pencil for solving this task, all subjects use the spatial strategy (Newton & Roberts, in press). When the change in task requirements is stable then high spatial ability subjects will change to a more appropriate strategy. When the task requirements cannot be predicted in advance, as with the sentence-picture verification task, a strategy that levels out the effort required for harder problems and easier problems is used by the high spatial ability subjects.

These intriguing studies with the compass directions task support an analysis of strategy change in terms of developing an appropriate level of abstraction in the representation. Analysing strategies in this framework indicates the commonalities between all three reasoning tasks considered here. Furthermore, spatial ability can be seen to be testing the same ability to develop strategies that use more appropriate levels of abstraction across these different tasks. Spatial ability is as much an indicator of strategy selection as it is of ability to use a strategy once selected. It is this issue of strategy selection that suggests that spatial ability measures may assess a style of information representation as much as an ability to represent. All the reasoning tasks considered present with strategies that are more or less appropriate for solving the problems, so selection of a level of abstraction for solving the task is confounded with ease of solution. Finding a reasoning domain where strategies using concrete and abstract representations are equally effective
will indicate whether spatial ability, for example, relates to a style for using certain strategies, or an ability to select an appropriate representation. This issue will be pursued in Chapters 3 and 4. In the meantime, understanding the process of strategy selection does not stop with linking it to measures of spatial ability. In order to approach an explanation, exactly what spatial ability measures are testing must be considered in more detail.

Roberts, Gilmore and Wood's (1997) explanation for why high spatial ability subjects are more likely to use the cancellation strategy hinges on the greater ability of these subjects to use the spatial strategy accurately which provides more likelihood that, for these subjects, they notice redundancies in the method and this leads to the development of the cancellation strategy. Explanations in terms of cognitive style – that strategy choice is a matter of preference – are inadequate, as Newton and Roberts (in press) show: those using the spatial strategy do not have any choice about this. Furthermore, explanations in terms of knowledge do not account for all the data: when provided with pencil and paper those previously using the cancellation method revert to the spatial strategy (Roberts, Gilmore & Wood, 1997). For task variants, subjects who have higher spatial ability are more likely to use the spatial strategy on the two-person compass directions task. Low spatial ability subjects continue to uncritically use the cancellation task after being shown this for the one-person version of the task. Furthermore, Macleod et al.'s (1978) high spatial ability subjects use a flat strategy which seems to be more efficient for harder problems, though there is extra expense for easier problems. Hence, spatial ability seems to lend an ability to selectively apply strategies as appropriate to the constraints of the task. Crowley, Shrager and Siegler (1997) suggest that the higher ability of these subjects frees up metacognitive resources which can be used to critically assess strategies for a given task.

All these accounts are given without regard to subject's performance on spatial ability measures. The next section discusses the sensitivity of different measures of spatial ability to strategic variation. This strategic variation can again be characterised in terms of level of abstraction used in the representation. Those with high spatial ability, the subjects that demonstrate strategic flexibility in the reasoning tasks, are those that use a variety of strategies on the items in the spatial ability measures. However, the link remains contentful if use of abstract or concrete representations is considered as a matter of style rather than ability, a perspective later offered by relating different representation use to use of alternative algorithms. The styles of processing with abstract or concrete representations will be considered further in Chapter 3.
2.4 Spatial ability: what does it measure?

Spatial ability can be defined as the ability to generate, retain, and transform abstract visual images (from Lohman, 1979, reported in Kyllonen, Lohman & Snow, 1984). There are numerous psychometric tests designed to measure spatial ability, and these have developed against a theoretical background where subfactors of spatial ability have been discovered. Hence, spatial operations, spatial visualisation, and spatial relations abilities have been distinguished as separate abilities measured by different tasks (Poltrock & Brown, 1984). However, fine-grained distinctions concerning the definition and subfactors of spatial ability tend to ignore the susceptibility of these tests to different strategic approaches to solving the problems. Indeed factor analyses will gloss over individual differences resulting from the application of more than one strategy. Many spatial tasks can be solved using representations that have been described as more "linguistic" than "spatial", and use of strategies that employ these "non-spatial" representations have been observed on many of the tasks designed to assess spatial processing. These contrasting strategies have often been described as "analytic" and "holistic", where holistic strategies preserve the unity of the presented stimulus, and analytic strategies attempt to decompose the stimulus into components which are then operated upon.

Lohman and Kyllonen (1983) reviewed studies where subjects were asked to introspect on how they solved spatial tasks. The results indicated that strategies differ between subjects both within a task and between different tasks. Lohman and Kyllonen note a general trend in strategic approach to spatial tasks where, as items become more difficult, there is an increase in the tendency to use what they term "analytic", or "non-spatial" strategies. Thus, a general review suggests that spatial task solution is replete with strategic influences. Several studies are reviewed in more detail below.

![Figure 2.3: Example of an item from the PFT.](image)

Snow (1980) recorded subjects' strategies on the paper folding test (French, Ekstrom & Price, 1968), or PFT. This task requires the subject to imagine the array of holes resulting from a piece of paper being folded then having a hole punched in it, and match this target to one of a set of multiple choice possibilities. An example item is shown in Figure 2.3, where the left three
diagrams show the folds, and the right five diagrams show the possibilities for the outcome of the hole-punching. Snow noted two strategies for examining the target figure: a subject could “mentally construct” the target from the folds, or alternatively the subject could use a “feature-extraction” strategy, where the subject attended to symmetric folds and number of folds in order to devise an encoding of the stimulus. For example, as the first fold is down the left-right diagonal in Figure 2.3, the resulting holes must be symmetrical about this axis. For examination of the multiple-choice items, the subjects used either a template-matching strategy, or a distractor-elimination strategy. The “mental construction” strategy specifies all the information in the array, whereas the “feature-extraction” strategy only specifies enough information to select the correct answer, abstracting away from exact position of holes and considering the arrangement of holes instead. The more concrete representation is holistic, as defined above, and the more abstract representation is analytic.

Cooper (1980) also found different strategic approaches to spatial tasks. She found that there were two distinct strategies used by subjects for a mental rotation and visual comparison task. Some subjects used a holistic approach, where the “visual memory representation” was compared in parallel with a test shape. Others used a more sequential, analytic comparison process, where features of the stimuli were compared serially. As with the PFT study, the analytic strategy employs a level of abstraction greater than that used in the holistic strategy, as only certain features of the stimulus are picked out for comparison with other features left unspecified. A similar pattern of strategic differences were found in a study by Kail, Carter and Pellegrino (1979) assessing performance on a mental rotation task. The Spatial Relations Test from Thurstone’s (1957) Primary Mental Abilities set of tests presents a letter-like stimulus with six rotated versions of the target, some of which are identical to the stimulus after rotation. The subject’s task is to say which ones are identical. They computed two latency measures: one to account for the time taken to encode the stimuli and make a response, and one time for the mental rotation. They found that the slope of the line for degree of rotation and time was widely different for some subjects, who seemed to take twice as long to rotate each stimulus as the modal group. They suggest that this is due to this group of subjects using an analytic method for rotating: they do not rotate the whole figure, but rotate parts of the stimuli sequentially. Rotating only a part of the stimulus demonstrates use of a strategy employing representations with more abstraction as elements of the shape are left unspecified.
Carpenter and Just (1979, reported in Kyllonen, Lohman & Snow, 1984) performed a study of a cube comparison task, measuring errors, latencies, eye movements and self reports, which revealed two distinct strategies suggestive of the holistic-analytic strategic difference. Some subjects represent and process the cubes holistically, other subjects process the stimulus sequentially, using a feature by feature transformation strategy. Again, the analytic strategy uses representations with more abstraction.

These studies strongly suggest the need for considering different strategies in terms of the use of different levels of abstraction in representation as subjects solve spatial ability tests. However, none of these studies provide data about the development of strategies as subjects learn to solve spatial ability tests. Lohman and Kyllonen (1983), however, developed a study expressly designed to address issues of strategy change and variation in a variety of spatial tasks. They gave their subjects mental synthesis tasks that varied in complexity. These problems required subjects to add simple shapes together and then match the result to a target which may be presented in a rotated position. They found that individual differences emerged at various stages. For memorising the initial stimuli in order to perform rotation or synthesis operations, three different strategies were evident. Subjects either stored each figure as a set of basic features, or decomposed the stimulus into simpler units such as triangles or rectangles, or attempted to apply a descriptive label to the stimulus as a whole. These three strategies vary in terms of the level of abstraction involved in the representation: decomposing the figure means more information is left unspecified and so is more abstract.

Some subjects responded flexibly by labelling some figures, and decomposing others, according to whichever strategy was easier. Subjects who used the decomposition strategy were faster at solving the task, and scored higher on other spatial tests. In terms of strategy change, as the task progressed, more than one third of subjects changed from a feature analysis to a decomposition strategy. No subjects showed the opposite direction of strategy change. This suggests that the most appropriate level of abstraction for the task is employed in the decomposition strategy. This hints at spatial ability measures being prone to strategy variation and appropriateness of strategy selection along precisely the same lines as those they reflect in the reasoning tasks. For the synthesis stage, subjects who scored highest on average on other spatial ability measures were those that exhibited the greatest flexibility in strategy for synthesis. For a task where three shapes had to be synthesised, these high ability subjects attempted to synthesise all three figures,
but if this was too complex, then subjects would recover and attempt to synthesise the first two figures, and if this failed, they would try to synthesise the last two figures. If all synthesis efforts failed, the three figures would be stored separately.

These studies all point towards the necessity of considering strategies in even the simplest spatial task. Alternative methods other than those purportedly tested by the tasks can be brought to bear on the various tasks. Many studies classify contrasting strategies as being either holistic or analytic, where holistic strategies use less abstraction in the representation than the analytic strategies.

Is it possible to say whether analytic strategies are more effective than spatial strategies for solving spatial tasks? The results of spatial ability studies are mixed. Lohman and Kyllonen (1983) suggest that analytic strategies are more effective, whereas Carpenter and Just (1979) found that the holistic strategy produced more accurate performance. Snow (1980) found that a strategy that combined holistic and analytic processes was most effective. Carpenter and Just (1979) found that use of the analytic strategy on the cube comparison task corresponded to poorer performance. Yallow and Webb (1977) found that low ability subjects on a range of spatial tasks reported using more verbalisations and spent more time evaluating and eliminating alternatives, whereas high ability subjects tended to know the answer before examining alternatives. The effectiveness of a strategy must depend on the task requirements. Lohman and Kyllonen's (1983) study suggests that for some items in the mental synthesis task analytic strategies were better, and for other items a holistic strategy is more appropriate. The better subjects are those that demonstrate more flexibility in selecting the appropriate strategy. Cooper (1976) found that holistic strategies were more effective for making "same" judgements for rotated figures, whereas analytic strategies were more effective for making "different" judgements.

In a study of the way individuals shifted strategy on a spatial task, Barratt (1953) found that for easy items in the DAT Spatial Relations test a holistic strategy was modal. For harder items, however, subjects tended to use an analytic strategy where no folding or unfolding of the stimulus pattern or response figures occurred, rather cues such as angle intersections were attended to. This variation in strategy is very like that observed by Snow (1980) in the PFT. A third strategy was when subjects started with the alternatives first, and then looked at the stimulus figures for matching, but the numbers of subjects using this strategy did not vary from easier to harder items. Myers (1958) reports introspective accounts of subjects solving a number of spatial tasks,
and notes a similar change from holistic to analytic strategies as problems get more difficult.

The strategic variation observed in these spatial ability measurement tasks suggests that what they test is not reducible to a single dimension. The generality of strategic variability in a range of tasks can be captured by distinctions in terms of holistic/analytic strategies. The differences between these strategies is in terms of the level of abstraction that they employ in the representations used.

Different tasks are more appropriate for one or the other strategy, but there is a general trend towards changing from using holistic to analytic strategies when the problems are harder. "Harder" can be defined in terms of the number and type of operations required to produce the unified representation, and in terms of the amount of information specified in the resulting representation. For harder problems the number of operations and the consequent load is greater, so more abstract strategies will reduce this load. However, using a more abstract strategy may require a degree of reprocessing during the task. If the wrong feature of the spatial stimulus is rotated then another may have to be selected. In the sentence picture verification task, when sentences contain negation operations are more complex on the verbal representations. This suggests that the types of process required by the holistic and the analytic strategies are different, and thus the underlying algorithms will be different. The computational differences in operating on abstract and concrete representations will be returned to in Chapter 4, but they remain in terms of the number and type of operation that has to be performed on the data structure.

Subjects that score highest on spatial ability measures are flexible in the strategy they use. This suggests that the flexibility of representation use in reasoning tasks by high spatial ability subjects may be a correlation of approaches rather than anything more contentful. However, there is evidence of styles at play in spatial task solution: there are subjects who utilise only an analytic strategy uncritically as well as those who use only a holistic strategy for a task. This point is worth pursuing, and the next chapter reviews several potentially useful measures for highlighting these issues of style and ability in representation selection.

2.5 Summary of Chapter

This chapter has made the case for the ubiquity of different strategies in reasoning tasks, and particularly the widespread evidence for what have been termed spatial and verbal strategies
used for the same task. The distinction between different strategies has been redescribed in terms of the level of abstraction present in the representations used, and this forges a qualitative link between verbal and analytic strategies, and between spatial and holistic strategies, in that the former tend to use more abstraction than the latter.

Studies relating spatial ability measures to reasoning performance show that high spatial ability subjects develop strategies with an optimal level of abstraction sooner. Explanatory accounts that invoke spatial ability must further probe the mechanism of strategy selection that spatial ability tests measure. The commonalities between strategy use on reasoning tasks and on spatial ability test problems are in terms of the extent to which the strategies use abstract or concrete information in the representations. This relates to the expressivity of the representational systems from which the representations are drawn. The type of operations that apply to the representations is going to be determined by the expressivity of the representational system, and the extent to which these operations are a matter of style or ability is, as yet, an open question.

The reasoning domains have presented tasks where different levels of abstraction are more or less appropriate for solving the task in terms of the effort or economy of using the representation. This makes styles and abilities hard to distinguish for these tasks, though the range of strategy types coupled with the different methods for solving spatial ability test problems suggests that some dimension of style is at play. The next chapter discusses some potential measures of preferences and ability for representing information with different levels of abstraction. It also presents some requirements on a reasoning domain where issues of style and ability can be distinguished.
Chapter 3

Selecting tests and tasks

Chapter 2 made a case for the widespread variety of strategies brought to bear on reasoning tasks and psychometric tests for spatial ability. These strategies exhibited some generality, however, in terms of the level of abstraction used in the representations. The first part of this chapter considers a number of conventional dimensions from the psychometric literature on styles and abilities that may help in assessing students' propensities and preferences for using representations with different levels of abstraction. In consequence, this will reflect abilities and styles for the different types of process correlated with the alternative representations.

The previous chapter illustrated that certain tasks are more effectively solved by constructing a more concrete representation, whereas others are better solved by using abstract representations. This warns of the need to "know thy task" as a pre-Socratic prerequisite before one can "know thyself". Ideally, a reasoning domain for study ought to be open to strategic variation, provide the potential for the use of different representations, and also the representations used by different strategies ought to be encapsulated by quantifiable differences in the abstractness of the representations used. Only then can the measures discussed in the first part of this chapter be used to distinguish preferences and abilities. Hyperproof, a multimodal, computer-based logic course, presents itself as a prime candidate for such a domain. The second part of the chapter illustrates the richness of Hyperproof as a didactic tool as well as a microscope for strategic variation. Some novel investigations of individual differences in strategy use in Hyperproof will highlight the processing distinctions between using strategies that involve abstract and concrete representations. Where there are these differences in representation, strategic and algorithmic descriptions
of individual differences prove to be complementary.

3.1 Strategies and Styles

The previous chapter discussed differences in strategy use for reasoning that have conventionally focused on alternative uses of representations. Frequently used candidates for reflecting different strategies have been spatial ability tests. Use of these tests to illuminate the selection and use of different strategies was shown to be in difficulty as measures of ability because the tests are prone to strategic variation. In addition, the tests are contentious as measures of strategy preference because the effective use of different strategies varied according to the task.

What would be useful for assessing strategy use in reasoning are measures of different “cognitive styles” (see Chapter 1). Five potentially useful measures from the psychometric literature are now discussed. These measures are variously referred to as “preferences”, “learning styles”, or “thinking styles”. Use of the term “cognitive styles” presupposes that preferences are due to differences in cognitive processes, making the assumption that learning is a mirror for cognition means that the terminological differences are due to the particular perspectives taken by the originators of the methods.

3.1.1 Visual/verbal preference

Paivio (1971), following in the imaginal footsteps of Galton (1883), attempted to assess what he termed verbal and nonverbal “ways of thinking”. These questions covered preferences for visual and verbal presentations of information. Example items questioned the subject’s preference for doing work that requires the use of words, or the ease with which the subject can generate a “mental picture” of a friend’s face. Such measures are pertinent to issues of level of abstraction in representation, as the visual/verbal distinction can be recast in terms of specific/expressive systems of representation. Gauging preference for visual/verbal information relates to assessing the subject’s preference for more expressive or more specific systems for representing. Paivio and Harshman (1983) found that responses to these questions were well-fit by a two factor solution, one relating to verbal preference, and one to nonverbal preference. Richardson (1977) took 15 items from Paivio’s “ways of thinking” questionnaire. This questionnaire is known as the
Visualizer-Verbalizer Questionnaire (VVQ). A factor analysis of this questionnaire performed by Kirby, Moore and Schofield (1988) found that the nonverbal dimension confounded an “imagery” dimension with a “dream” dimension (the vividness and frequency of dreams experienced by the subject). They then constructed a 30 item questionnaire that assessed preference for verbal and visual information separately from that of dream vividness.

Kirby et al. tested whether the learning styles of their questionnaire were related to ability measures. Their verbal ability measure was the ACER Higher Test Form ML from the Australian Council for Educational Research (1981). This consisted of vocabulary, verbal similarities, verbal reasoning, and verbal analogy problems. There were two spatial ability measures: the Card Rotations test, and the Surface Development test (both from Ekstrom, French & Price, 1963). They found that verbal preference corresponded with scores on the verbal tests ($r = 0.32, p<0.005$), and visual preference correlated with the Surface Development test ($r = 0.27, p<0.005$). They did not find a significant correlation between visual preference and the Card Rotations task. But, they also found that verbal preference correlated with the Surface development test ($r = 0.17, p<0.05$), and that visual preference correlated with verbal test score ($r = 0.32, p<0.005$). Hence, there is no neat connection between score on verbal tests and verbal preference or between score on spatial ability tests and visual preference.

The lack of systematic connections between visual/verbal style and performance on verbally-based or spatially-based tasks argues against the use of such measures, as does their derivation from introspective accounts. For these reasons, the VVQ is not considered for further use.

A further criticism of visual/verbal preference as a measure of strategy differences comes from a study by Oberlander, Cox, Monaghan, Stenning and Tobin (1996). They showed that differences in reasoning with combinations of graphical situations and expressions in propositional calculus are not due to differences in the use of either the visual or verbal modality. Rather differences in strategy are better expressed in terms of the degree of abstraction employed in representations. In the current context, the visual-verbal preference can be subsumed under analyses that explore expressivity/specificity in reasoning tasks.

Paivio (1986) turned to focus on the processes induced by verbal and spatial representations, rather than properties of visual/verbal information. His “dual-coding” theory assumes separate systems for dealing with verbal and nonverbal information. The properties of the processes within each system are resonant of previous discussions of the different properties of strategies:
The verbal system is specialized for sequential processing whereas the nonverbal system is specialized for synchronous or parallel processing of multiple representational units. (Paivio, 1986, p.71).

This shift in emphasis towards process pre-empts the holistic/analytic style which attempts to characterise the difference between “sequential” and “parallel” processing. Holistic/analytic style is the next dimension to be considered.

### 3.1.2 Holistic/analytic style

Anderson speaks of cognitive psychology's “almost fatal attraction to bold, general claims about human cognition” (Anderson 1993, p.2). The distinction between holistic and analytic processing is a casualty of such boldness and generality, as it has been used to distinguish binary performance on many dimensions in both social psychology and cognitive psychology. The following discussion presents studies that contribute to a coherent conception of the holistic/analytic distinction that enables its connection to reasoning, even though their starting points are as diverse as hemispheric asymmetry, concept formation, and geometry problem solving.

The view that the difference between holistic/analytic processing underlies the visual/verbal distinction is pursued by Bradshaw and Nettleton (1981) in their review of the holistic/analytic distinction relating to hemispheric asymmetries, previously considered as a verbal/spatial difference. Initial claims about hemispheric specialisation from a functional perspective expressed the distinction in terms of a verbal advantage for the left-hemisphere and a non-verbal/spatial advantage for the right hemisphere. Broca and Wernicke have both claimed a verbal role for the left hemisphere, finding as they did that certain areas of the brain that are essential for spoken language production are located in the left temporal lobe. Studies on normal subjects employing a variety of lateralised presentation methods have found that the left hemisphere is advantaged for verbal presentations, whereas the right hemisphere is advantaged for non-verbal tasks. In the auditory modality, Goodglass and Calderon (1977) have found a right ear advantage for verbal material such as backwards speech, nonsense syllables, and words and sentences maintaining normal syntactic structure. In the visual modality, a right visual field advantage emerges for verbal stimuli, but a left visual field advantage has been found for visual materials where the principle feature of the task is colour discrimination (Davidoff, 1976), dot localisation (Bryden, 1976),
the perception of line orientation (Atkinson & Egeth, 1973), or curvature (Longden, Ellis & Iversen, 1976), as well as a host of other non-verbal features (for a review see Bradshaw & Nettleton, 1981).

As Bogen (1969, 1975) observes, the mode of processing seems to be more fundamental than the nature of the preferred stimuli in hemispheric specialisation: the left hemisphere deals with analytic processes, the right hemisphere is concerned with holistic processing. This reflects the distinction that Paivio makes with regard to spatial/verbal information: the processes operating on the stimuli are more fundamental than the stimuli themselves. Thus, the right ear advantage for verbal materials disappears when the task demands are altered. Bever, Hurtig and Handel (1976) attempted to demonstrate that left hemisphere specialisation is for analytic processing rather than a specialisation for verbal material. When the stimuli are presented to one ear, there is a right ear advantage for the detection of an initial phoneme of a consonant-vowel-consonant syllable, but there are no ear differences in the detection of the whole syllable as a target, though other studies have indicated a right ear advantage for syllable identification (Bradshaw & Nettleton, 1981, p.57). Nebes (1978), in a similar vein, illustrates that the right hemisphere is superior at perceiving the relationship between component parts and the whole configuration, and in performing spatial transformations of the visual input. He posits that the right hemisphere forms a spatial and cognitive map of our surroundings from incomplete sensory information, basing this opinion on the right hemisphere being advantaged when forming a complete Gestalt (e.g., a circle) from incomplete information (e.g., arcs of a circle), while on the other "hand", the ability to discover or isolate a shape within an irrelevant background is a function of the left hemisphere.

Bradshaw and Nettleton (1981) summarise their review of the literature by noting there is a left hemisphere specialisation for temporal order, sequencing and segmentation. This is based on studies that have found a right ear advantage for complex, rhythmic, sequential series of acoustic stimuli, and a left ear advantage for simple, discrete, unidimensional, global stimuli. Also on evidence for a right-body side tactual advantage for sequentially ordered stimulation, and a left-side advantage for stimuli perceived in terms of spatial distribution. Thus, distinctions have moved away from a verbal/non-verbal distinction towards embracing a distinction in terms of the holistic/analytic style of processing.

Kemler Nelson (1984) uses the holistic/analytic distinction to describe different processes leading to concept formation:
On the analytic mode, stimuli are compared and contrasted according to their constituent properties or attributes; in the non-analytic or holistic mode, they are related according to global relations of overall similarity (p.735).

These different approaches to concept formation can be tested by requiring the subject to group together stimuli that vary along two perceptual dimensions (Garner, 1974). Figure 3.1 shows a set of wedges that vary in terms of size of angle and length of side (Smith & Baron, 1991). Subjects that group the left and middle stimuli together are driven by the similarity of the length of side of the two shapes, so their decision is dictated by judgment on a single dimension. Subjects that group the middle and right stimuli do so due to the "family resemblance" of the two properties: the two shapes do not have identical side length or angle, but are closer on both dimensions. Kemler Nelson and colleagues (Foard & Kemler Nelson, 1984; Kemler Nelson 1984; 1989; Kemler & Smith, 1979; Smith, 1989; Smith & Baron, 1981; Smith & Kemler, 1977; Smith & Kemler Nelson, 1984) have shown that children are more likely to be holistic, but adults more likely to be analytic. Adults under dual task conditions, or forced to make judgements at speed, are more likely to group stimuli holistically. This leads Kemler Nelson to contend that "holistic processing, guided by global similarity relations rather than abstracted stimulus properties, may be frequent, fundamental and primitive in human cognition" (Kemler Nelson, 1984, p.735).

Figure 3.1: Stimuli to test for holistic/analytic concept formation.

In a similar vein, Schmeck and Geisler-Brenstein (1989) noted that holistic (or global) learning processes seem to involve habitual preference for "a broad focus of attention, formation of impressions, noticing similarities, more interest in wholes than in component parts, and preferences for more random, less orderly presentations of information. Analytic processing involves a narrower attentional focus, retention of facts and details, noticing differences, more interest in parts than wholes, and preferences for ordered (usually sequential) presentations of information" (p.114). They go on to discuss this issue of cognitive process with respect to development, stating that "the highest levels of development require integration, specifically integration of the earlier more childlike and random, global functioning with the newer self-controlled, analytic skill" (p.116). This suggests that the terminological similarities with the hemispheric asymme-
Attempts to locate holistic/analytic processes in complex tasks have been conducted. Boulter and Kirby (1994) suggest that strategies used to solve transformational geometry problems can be usefully described in terms of holistic/analytic processes. The holistic strategy involves manipulating the whole object, focusing on the whole rather than breaking it down into parts or features. In the analytic strategy, the subject considers distinct features or parts of the problem or display in question when attempting to arrive at a solution. With respect to transformational geometry, a strategy that involved the shapes being analysed as being made up of parts, and then moving the figure by "pieces" was described as an analytic strategy. A strategy where the shape was moved as a whole was described as a holist strategy. Results of the study showed that some questions were best solved by one or other strategy - two questions were more successfully solved by students using an analytic strategy, and one question was best solved with a holist strategy. Therefore, unsurprisingly perhaps, the most successful approach to solving the range of problems seemed to be flexible strategy deployment.

Some studies have attempted to relate the holistic/analytic distinction to complex problem solving. The difficulties of attaining an agreed conception of the relation between these processes and complex tasks is evident in such work. Alesandrini, Langstaff and Wittrock (1984) measured holistic and analytic abilities in both pictorial and verbal modalities. Their approach was in terms of ability rather than style, so their studies are indicative of performance on tests that prioritise one form of processing over another. Tests that require the subject to focus more on the parts than on the whole or to pick out the parts from the whole in order to get the correct answer were taken to be analytic tasks. Holistic tests were classified as those that require the subject to mentally combine parts into a whole or to focus more on the whole than on the parts in order to get the correct answer (p.152). To measure pictorial-analytic ability, Witkin, Oltman, Raskin and Karp’s (1971) Group Embedded Figures Test (see discussion of field-independence/dependence below) was used “because it requires the respondent to analyze a spatial whole in order to pick out a component part” (p.153). For pictorial-holistic ability, the Street Figure-Completion test was used (Street, 1931). This “requires the respondent to recognize a spatial whole from the several component parts shown” (ibid.). Verbal-analytic ability was measured with the Nonsense Syllogisms test (French, Ekstrom and Price, 1963): “since the respondent must sequentially focus on pieces of information to derive another piece of information, this test was selected as a measure of verbal-
analytic ability” (ibid.). Lastly, verbal-holistic ability was measured by the Similarities subtest of the Wechsler Adult Intelligence Scale (WAIS) (Wechsler 1958), where the subject is required “to see a higher level of relationships between the paired objects. If the subject gives the higher level commonality of the two objects, a higher score is given than if the response is analytical, focusing on specific attributes of each object in the pair” (ibid.). This latter test has also been used as a measure of verbal-analytic ability by a subset of the same authors, where higher scores are given for what are termed “symbolic-abstract similarities, and lower for spatial-relational similarities” (Wittrock & Alesandrini, 1990, pp.493-494). This test and the Street Figure-Completion test have (apparently) “been frequently used in neuropsychological studies to assess spatial-holistic and verbal-analytic abilities, respectively” (Wittrock & Alesandrini, 1990, p.494). Unfortunately, the only reference to use of these tests is the Street test used by Bogen et al. (1972).

There are evident difficulties in linking holistic/analytic processing to complex problem solving tasks. Using tests that are as complex as the task to which they are compared does not clarify the nature of preference for using different processes in reasoning tasks. Das, Kirby and Jarman (1975) have suggested that holistic/analytic processes are one of a family of related descriptions of different processes that are brought to bear on complex cognitive tasks. Clark and Frisby (1980, reported in Alesandrini, Langstaff & Wittrock, 1984) take a similar line, suggesting that the processes contribute to different ways of thinking about problems. Holist (or synthetic) thinking enables an individual to put the discrete parts of an information presentation together into a meaningful whole by focusing on the whole rather than on the parts. Analytic thinking entails critically categorising and classifying information.

After reviewing several studies for assessing holistic/analytic processing differences, Beyler and Schmeck (1992) suggest that a unified concept of holistic/analytic is emerging. It certainly appears as a relevant and important distinction for expressing the strategies observed in spatial ability tests, and thus for capturing the difference between use of different levels of abstraction in representation. The problem is what tests are to be used in order to assess it. Those used by Alesandrini et al. (1984) seem unsatisfactory as their complexity means that at best they can offer a correlation rather than an explanation. The concept formation test used by Kemler Nelson and colleagues is promising, but such measures seem very fragile, as responses change with slight changes to stimuli or to their presentation (Foard & Kemler-Nelson, 1984). Miller (1988; 1991) suggests that the holistic/analytic processing distinction underlies both the field-
dependence/independence and the serialist/holist styles. These measures are now considered as potential tools for assessing performance in complex tasks.

3.1.3 Serialist/holist style

In order to explore different learning strategies on an unfamiliar task, Scott devised an experiment known as the "martian animal classification task" (Pask & Scott, 1972). Students had to learn taxonomies for two different sorts of animal: the "clobbit" and the "gandelmuller". Information about the different species was presented on 40 cards which provided different levels of information. 10 cards were typical subspecies, 15 were of contextual data about habits, 5 were test types in the taxonomy, 7 were of physical characteristics, and 7 indicated why names of parts or behaviours are used. Subjects were free to select any card they wished, but were required to give a reason why they chose that particular card. Strategies were distinguished in terms of whether large global predicates were tested, or whether the subject preferred step-by-step approaches to learning. The holist learner prefers relations of topics, preferring to structure their learning at the global level, and investigating individual cases or cards presenting information at a lower relational level to fill out this structure. The serialist learner builds up a structure by investigating particular instances of the animals, creating abstractions from these individual cases. This distinction suggests similarities with the holistic/analytic processing distinction, but applies these concepts to learning situations. Pask expresses the distinction with relation to learning tasks in general: "Serial learners showed intention to search for specific data. Holist learners showed intention to test a large predicate or relational hypothesis" (Pask, 1988, p.90).

Alternative formulations of the serialist-holist learning strategy have been made in terms of the number of goals and working topics maintained during learning. The holist has many subgoals under his main topic, whereas the serialist has only one goal which may be the main topic. Once this goal is achieved, the serialist will move on to the next goal or topic (Pask, 1976). The distinction has also been supported with reference to questions asked by the student during "feedback" routines where the student recounts their knowledge of the domain. Holists ask questions about broad relations, to support their hypotheses about generalisations. Serialists' questions concern narrower relations, and hypotheses are specific. In the martian animal classification task, a holist asks questions such as "Are there more kinds of Gandler with mounds?" whereas a serialist questions a specific attribute: "Which kind of Gandlemuller has no sprongs...?" (Pask, 1976,
The serialist/holist distinction in strategy is described within a particular learning situation, so it is an expression of the particular relationship of the student to a particular task. Thus, strategies are mutually exclusive for a given learning situation. Certain tasks are likely to be better solved with one strategy over the other, but what in general disposes a student to use one or other of these learning strategies? The disposition a student has towards using a certain strategy defines the student's individual learning style. A student who uses a holist strategy frequently is classified as a holist, or comprehensionist, learner. The frequent user of a serialist strategy is named a serialist, or operationalist, learner. A comprehension learner readily picks up an overall picture of the learning situation, and recognises clearly where information can be obtained to fill out the details. Comprehension learners have the ability to build descriptions supplemented by associations between topics. Operation learners pick up rules, methods and details, but lack the awareness of how the elements of the learning situation fit together. The comprehension/operation distinction in learning style does not divide students into distinct classes – there is no mutual exclusivity as there is in defining the holist/serialist learning strategies. Instead, the scale is continuous, biases towards certain learning strategies are a matter of degree, and some students demonstrate an ability to effectively apply both serialist and holist strategies. These students are termed "versatile" learners.

However, students are often very inflexible with regard to use of a learning strategy, thus contextual effects of different tasks were found to be less a determiner of choice of strategy than stylistic consistency (Pask, 1975). Once a strategy has been chosen by a student it is rare that it is relinquished, even if it proves difficult to execute, resonating with Roberts, Gilmore and Wood's (1997) study on strategies for different versions of the compass directions task. Pask observes that only "strong advice" will encourage the student to start afresh. He goes on: "sharp strategic distinctions occur because students become locked into one strategy to the exclusion of others" (Pask, 1976, p.130). This means that the probability of a student using a particular strategy for a given task can be predicted in advance by measuring the student's bias towards a serialist or holist strategy for learning. Furthermore, task constraints indicate that every holistic or serialistic strategy is not efficient or effective. Thus, the student must be able to tailor their choice of strategy to the situation. Pask describes what he terms "learning pathologies" of students who approach every problem using the same strategy, no matter how inappropriate. Unsuccessful
application of a holist learning strategy is termed "globetrotting", whereas misapplication of a
serialist strategy is labelled "improvidence". Use of analogies during learning is an example of
these pathologies in action. The globetrotter will apply analogies, but inappropriate ones that re-
late to irrelevant or vacuous elements, or relate the wrong elements to the analogy. Improvidence
is characterised by the lack of application of analogies, or the failure to use a common principle to
relate elements (Pask, 1976, p.140). However, learning will be affected by interactions of presen-
tation style and learning style. Materials in the martian animal classification task were adapted to
favour a serialist or a holist strategy. Students presented with materials that matched their spon-
taneous approach to the task learned much better, requiring far fewer instructional presentations
before they learned the task. Other studies by Pask have shown that mismatched conditions
mean longer learning periods and information being retained for shorter periods. Pask (1976)
writes: "a mismatched condition leads to grossly inferior performance and a pronounced failure
to comprehend the principles underlying the subject matter" (p.132).

Pask is at pains to note that an effective serialist learner acquires the same understanding of
the learning situation, it is just their approach that is diverse. The serialist gains understanding
from particular instances, or a routine approach to learning. The holist learner gains insight from
working out the details of the situation from the global perspective, preferring freedom to move
between informational levels, filling out the details of a structure constructed at a more general
level. Pask contends that the end-product is no different, what varies is the point of focus during
learning.

In summary, Pask has shown that there are fundamentally different ways in which complex learn-
ing situations are approached by students. He has indicated that there are different ways of learn-
ing the same material, and these different approaches can be broadly characterised in terms of
whether the student treats the material as a whole, later filling out the structure of the situation
with detail, or whether the learner prefers to focus on the particulars of the presented material,
and gradually build up an overall sense of the situation. This characterisation of the learning
style suggests a difference in terms of the level of abstraction to be employed in the learner's rep-
resentations. The holist begins with an abstract, schematic structure, and fills out the details later,
whereas the serialist begins with more concrete representations, generating comprehension of
the whole subject by later abstracting over these cases. Pask has also shown that an individual's
learning style determines, for a given task, which approach will be taken towards the new learn-
ing situation. Though a student's approach to different situations can vary according to the task, most students are fairly inflexible in their approach to novel problems, demonstrating learning pathologies that limit their strategy application. Thus, the learning style of a subject is useful in depicting the process of problem solving.

Furthermore, the similarity of concepts between the serialist/holist learning style and the holistic/analytic processing distinction supports Miller's (1989; 1991) contention that the processing differences underly the learning styles. Pask considers the processes that may underly the style:

> ... holism and serialism appear to be extreme manifestations of more fundamental processes, which are induced by systematic enforcement of the requirement for understanding” (Pask, 1976, p.133).

The fundamental processes that the serialist/holist learning style taps into are the holistic/analytic differences that are inherent in the use of different levels of abstraction in representation. A small-scale study on University of Edinburgh students related serialist/holist preference as tested by Clarke's (1993) question (see Chapter 6 for further details of this question) to holistic/analytic processing using Smith and Baron's wedges stimuli (Figure 3.1). There was a suggestion that serialists made more dimensional, or analytic, judgments than holists (8.0 compared to 4.6 out of 32 items, \( t(10) = 1.69 \), one-tailed \( p=0.06 \)). Due to the variability of these concept formation measures, however, this was not pursued. A measure of serialist/holist style will be further considered as highlighting the essential differences between representation use in Hyperproof, considered later in this chapter, and a measure will also be used in the empirical study reported in Chapter 6.

### 3.1.4 Field-independence/dependence

Field-independence/dependence (FID) was also considered as a style which trades on the holistic/analytic processing difference, according to Miller (1989; 1991).

Gottschaldt (1926), in his studies of the roles of contextual factors and past experience in perception, constructed a test where a target shape was hidden in a complex line-drawing (summarised in Witkin, Oltman, Raskin & Karp, 1971). When the context plays a strong role in the subject's perception, the task was found to be more difficult as the disembedding of the shape is harder
for these individuals. The Embedded Figures Test (EFT) was derived by Witkin et al. (1971) from this test, which bears a strong resemblance to the Hidden Figures Test (HFT), devised by French, Ekstrom and Price (1963). An example item from the HFT is shown in Figure 3.2. The task in the HFT is to say which one of the five shapes shown at the top of the diagram is contained in the complex figure. Subjects demonstrate different abilities to locate the shape in the complex array of lines. Individuals that can perform this task quickly are more field-independent, as the context of the shape plays less of a role in their perception. Individuals that are slow on this task are more field-dependent, where their perceptual processes are much more determined by the context. This is the basis of FID as a style.

Another test which helped to define the FID dimension was the Rod Frame Task (RFT) where the subject is sat in a room, then the room is tilted, and the subject is instructed to move a rod until it is vertical. Subjects who are field-independent are better at this task than field-dependent subjects, reflecting an ability to ignore the immediate perceptual context, and prioritise vestibular over perceptual information. Witkin, Dyk, Faterson, Goodenough and Karp (1962) state the dimension as measured by the tasks thus:

[the task] requires the person to separate an item from the field or context of which it is a part and which therefore exerts a strong influence upon it; to "break up" a field or configuration. The person with a more field-independent way of perceiving tends to experience his surroundings analytically, with objects experienced as discrete from their backgrounds. The person with a more field-dependent way of perceiving tends to experience his surroundings in a relatively global fashion, passively conforming to the influence of the prevailing field or context.

The field-dependent subject perceives information in the field as "fused together", whereas field-independent individuals experience parts of the field discretely from the structure of the ground. This distinction between treating information globally or analytically led to a redescription of
the measure in terms of whether the subject has a “global” or “articulated” style of information processing.

Definitions of the FID dimension have therefore focused on two related competences. The first is the ability to discriminate parts of the field or break up an organised stimulus so as to separate parts of it (Witkin et al., 1971, p.5). This has been termed “analyticity”, and relates to Alesandrini, Langstaff and Wittrock’s (1984) use of the EFT to assess pictorial-analytic ability. It contrasts with “global” processing, where the embedding context is not overcome by the subject, but is an inherent part of their perception and problem solving approach. The second competence is in terms of applying structure to a field. Witkin et al. (1962) found that subjects scoring high on the EFT were more likely to impose structure on inkbolts on a page, even if the inkbolts were entirely random. Whereas the holist-serialist style assesses how the structure of the information is developed by the learner, the FID dimension analyses the structure itself. It is thus expected to be a measure complementary to the serialist-holist dimension. Indeed, Entwistle (1979) has claimed that field-independence co-occurs with serialist learning preferences. To assess similarities and differences between these measures, both are used in the empirical studies.

The FID dimension is classified as a style, rather than an ability measure in that subjects that score low on tests such as the HFT or the EFT are good at certain tasks. This reflects the distinction between styles and abilities as stated in Stenning and Monaghan (in press), where a certain style is laudable for one task and lamentable for another. Scoring low on an ability measure, on the other hand, is never characterised in an optimistic light. These differences in performance according to task requirements for FID styles are exemplified by the range of responses to learning programmes and reasoning domains discussed below. Some confounding of ability with style does take place, however, as Linn and Kylonen (1981) found correlations between EFT and the paper folding test ($r(60) = 0.56$). Witkin and Goodenough (1977, cited in Witkin et al., 1971) admitted that spatial ability and the EFT score overlap. Poltrock and Brown (1984) go so far as to use the EFT as a measure of what they term “spatial visualisation” ability (one of the subfactors of spatial ability that they propose). Such difficulties in discriminating style and ability led Cronbach (1984) to conclude that “there seems to be no point in separating [FID] from the concept of fluid ability” and that “it is a variable to match on, not to sort cases on” (p.267). However, spatial ability measures have been shown to be vulnerable to strategic variation, and so it may be that ability measures demonstrate style differences as much as style measures contain ability factors.
Such difficulties advocate for careful assessment of task requirements in order to apply style or ability measures in an explanatory framework.

The FID dimension has been associated with complex problem solving in a number of different domains. Most of these concentrate on establishing the connection between field-independence and scoring well on reasoning tasks. For example, the EFT correlates strongly with Raven's progressive matrices (Linn & Kyllonen, 1981), and is strongly associated with a factor that maps onto Fluid Intelligence (Cattell, 1963). Heller (1982) found that field-independent students were better at solving mathematical or science-based problems than field-dependent students. Davis and Haueisen (1976) found that field-independent subjects are more efficient at testing hypotheses when learning and solving set-membership problems. The field-independent subjects were better at reformulating hypotheses when the data set altered, field-dependent subjects maintained hypotheses that they used previously even if these were not appropriate in the changed context. Wise (1980) found that field-independent subjects are better at categorising visual stimuli, especially when the stimuli were complex.

Some studies have investigated the association between the FID dimension and strategy use. Pascual-Leone, Ammon, Goodman and Subleman (1978) relate the FID dimension to strategy selection. They found that field-independent subjects tended to select more appropriate strategies for solving problems. Field-dependence relates to an inability to critically assess and adapt ineffective strategies. Phifer (1983) found that field-dependent students tended to use a single strategy for a text comprehension task, whereas field-independent students used several. Furthermore, the field-independent students tended to fit their strategies to the style of material being presented. If the material was "social", they attempted an overview of the information, if the material was "scientific" in content then more attention was paid to terminology. When mathematics teaching material was highly structured, this benefit field-dependent students, whereas minimal guidance was most effective for teaching field-independent students (Adams & McLeod, 1979).

The FID dimension presents itself as a potential candidate for investigating issues of representation and strategy in reasoning due to several qualities. It has been shown to relate to different types of information processing: whether information is treated as "fused" or composed of analysable parts. This suggests that field-dependent students would be less likely to decompose a stimulus in a spatial ability task, for example, and thus less likely to utilise different levels of
abstraction in a task, whereas field-independent students may be more inclined to use an analytic, or abstract representation, approach in their problem solving. Also, though the FID is given a moral gloss in several of the studies on problem solving, as with the serialist/holist style it remains a measure of style rather than ability, enabling a disentangling of these issues in reasoning task performance.

3.1.5 Abstraction-concreteness

The previous dimensions considered have attempted to assess the different processes relating to use of representations that vary in terms of the level of abstraction that they employ. However, use of different levels of abstraction has been examined directly as an individual differences dimension. Bruner (1957) and Harvey, Hunt and Schroder (1961) have described human development and learning in such terms. Their conception of intellectual and societal advancement is in terms of a move from concrete to abstract understanding, which reflects Kemler Nelson's (1984; 1988) discussion of development from holistic to analytic processes in concept formation.

Harvey, Hunt and Schroder (1961) contend that human development and evolution can be characterised in terms of a change from concreteness to abstraction in information representation. More concrete information tallies with a structure restricted to, or dependent upon, the physical attributes of the activating stimulus (p.3). They relate this distinction to an earlier formulation by Bruner (1957) that abstracting means "going beyond the information given" and that personal development is a progression towards greater abstraction. Children's development on the number conservation task is a clear example of this, where the presentation is singularly attended to by young children, but abstraction occurs developmentally as quantity is conceived to be independent of form (Piaget & Inhelder, 1958). Furthermore, Harvey, Hunt and Schroder view development not just as a move towards more abstract representations, but towards a greater flexibility of approach to problems that accompanies the ability to abstract from stimuli in the environment. Variation in the level of concreteness-abstractness results in differences in "stimulus-boundedness", which is the extent to which the receiving and responding individual is restricted to or can go beyond the physical characteristics of the immediately impinging stimuli in organising his evaluation and experience of a situation. The greater the individual's ability to abstract, (1) the greater the ability to transcend immediacy and to move more into the temporally and spatially remote, and (2) the greater the capability for abstracting relationships from objects of
experience and organising them in terms of their inter-relatedness.

Goldstein and Scheerer (1941) have noted that the ability to abstract from the environment provides the ability to verbalise accounts of experience, to shift reflectively from one aspect of the situation to another; to hold in mind simultaneously various aspects; to grasp the essential of a given whole; to break up a given whole into parts, to isolate and to synthesise them; and to abstract common properties reflectively, and to form hierarchic constructs (p.4). They constructed tests to assess a subject’s “ability to assume the abstract attitude”, but these tests were designed for patients in a clinical setting who appeared to be pathologically confined to the “concrete”, as a result of traumatic brain injury¹. Such measures, then, are perhaps not appropriate for assessing performance of a normal student population.

Gregorc (1984) distinguishes subjects' predispositions to perceive information either abstractly or concretely, and operate on this information either sequentially or randomly. Abstract perception of information is the ability to process using reason and intuition, contrasting with concrete information processing which refers to the physical characteristics of input to the physical senses. Sequential processing style means that the subject has a preference for organising information in a linear, step-by-step manner, whereas random processing creates a network of information where data is related in various ways.

Stanovich (1999; Stanovich & West, in press) locates the ability to represent information abstractly as the greatest determiner to success in education and the modern workplace. The complexity of life in a technological age imposes demands such that operating on abstract information is necessary for success (Gottfredson, 1997). Holistic/analytic processes emerge as indicative of the use of abstract and concrete information in discussions of the properties and consequences of working with these different types of information. This supports the use of FID and serialist/holist measures for assessing preferences for using different levels of abstraction in representing problem information.

¹It is equally possible that patients may demonstrate pathological abstraction. For example, cases of prosopagnosia where individual face-recognition is impaired even though, in general, faces are still recognised as such (Sacks, 1985).
3.1.6 The GRE test

The GRE test is included here as a measure of style, as it has been found to relate to variations in the use of external graphical and verbal representations to support reasoning. Furthermore, differences in the use of graphical abstraction have been observed as relating to this test. A fuller consideration of the relationship between the GRE test and use of different levels of abstraction in representation will be given later in this chapter in connection to Hyperproof.

The GRE analytic reasoning test (hereafter, the GRE) presents itself both as a suitable task for studying individual differences in reasoning, as well as a potentially useful measure of stylistic variation in problem solving in other reasoning domains. It was originally intended as a measure to predict students’ success at graduate school, introduced in 1977 as a response to requests by universities to include a measure of abstract reasoning ability in the GRE. For a comprehensive history and review of the analytic reasoning test’s conception and development, see Cox (1996).

The GRE test distinguishes two subscales on the test: analytic reasoning and logical reasoning. Examples of each of these are shown in Figures 3.3 and 3.4, respectively. Analytic reasoning items "are usually constraint satisfaction puzzles for which diagrams are often useful" (Stenning, Cox & Oberlander, 1995, p.6), whereas the logical reasoning problems "involve argument analysis, a kind of verbal reasoning problem" (ibid.).

Analytic problems constrain the number of situations consistent with the given information. The problem shown in Figure 3.3, for example, is consistent with only one model: office assignment is determined by the information. This is a reason why many students attempt this sort of problem by constructing a diagram: concretising information is optimal for solving this sort of problem. Logical reasoning problems do not offer such constraints on the situations consistent with the information: they are indeterminate and must therefore be solved using a more expressive system, so representations will be more abstract. Students tend to solve this sort of problem by using verbal external representations (if they use any at all), for example underlining key words, or providing verbal summaries of key points of the arguments.

The analytic reasoning subscale of the GRE (GREA) has been shown to be indicative of stylistic preference in reasoning, rather than being a “reasoning ability” measure. Students who score well (GREA-Hi) on the GREA and those that score less well (GREA-Lo) learn differently from logic courses depending on whether the course has a graphical component, as with Hyperproof, or whether the course is traditionally syntactic. GREA-Hi students learn better from a graphical
An office manager must assign offices to six staff members. The available offices, numbered 1-6 consecutively, are arranged in a row, and are separated only by 6-foot-high dividers. Therefore, voices, sounds, and cigarette smoke readily pass from each office to those on either side.

Miss Braun's work requires her to speak on the telephone frequently throughout the day.

Mr White and Mr Black often talk to one another in their work, and prefer to have adjacent offices.

Miss Green the senior employee, is entitled to Office 5, which has the largest window.

Mr Parker needs silence in the office(s) adjacent to his own.

Mr Allen, Mr Parker and Mr White all smoke.

Miss Green is allergic to tobacco smoke and must have non-smokers in the office(s) adjacent to her own.

1. The best location for Mr White is in Office 1, 2, 3, 4 or 6?

2. The best employee to occupy the office furthest from Mr Black would be Mr Allen; Miss Braun; Miss Green; Mr Parker or Mr White?

3. The three employees who smoke should be placed in Offices 1, 2 and 3; 1, 2 and 4; 1, 2 and 6; 2, 3 and 4; or 2, 3 and 6?

Figure 3.3: Analytic reasoning item in the GRE test.

Excessive amounts of mercury in drinking water, associated with certain types of industrial pollution, have been shown to cause Hobson's Disease. Island R has an economy based entirely on subsistence-level agriculture; modern industry of any kind is unknown. The inhabitants of Island R have an unusually high incidence of Hobson's Disease.

Which of the following can be validly inferred from the above statements?

- I Mercury in drinking water is actually perfectly safe.
- II Mercury in drinking water must have sources other than industrial pollution.
- III Hobson's Disease must have causes other than mercury in drinking water.

II only; III only; I or III but not both; II or III but not both; or II or III or both?

Figure 3.4: Verbal reasoning item in the GRE test.
course, GREA-Lo students learn better from the syntactic course. This result is returned to when Hyperproof is described later in this chapter. In addition, GREA-Hi students construct proofs that incorporate a greater degree of graphical abstraction than those produced by GREA-Lo students in the graphical logic course. In summary, then, students seem to respond differently to the level of abstraction used in their representations and this is captured by performance on the GREA test.

The GRE test provides two-for-the-price-of-one value in that it provides a domain where the effectiveness of concretising information is variable, though the presentation of problems is constant (they are all verbal, and all on the same test paper). Furthermore, this variability in effectiveness of different levels of abstraction in representation can be captured in terms of the “determinacy” of the problem. It also provides a diagnostic tool for the use students make of different levels of abstraction which can then be used to assess performance on other reasoning tasks. The next section highlights the potency of this abstraction-level identification facility by stipulating the requirements on appropriate reasoning tasks for considering issues of representational and strategic differences in problem solving.

3.2 Reasoning task requirements

The tasks discussed in the last chapter—linear syllogisms, the sentence-picture verification task, and the compass directions task—are all useful for exhibiting individual differences in reasoning in terms of differences in the representations used in strategies for reasoning. However, they are limited when it comes to providing computational descriptions of these strategic differences. This has the result that the effectiveness of different strategies as the task constraints change cannot easily be assessed. The requirements for a reasoning domain that can clarify these issues, and consequently address issues of style and ability for using information with different levels of abstraction are now listed.

The level of abstraction in the representations employed in the strategies previously reported for reasoning task performance are hard to assess. A domain where representations are externalised would assist in analysing the level of abstraction used by the student. The traditional focus on “output” of the task also means that use of different strategies are not always clearly determined by the data. For example, Sternberg and Weil (1980) had to look at which model fit their subjects...
best, rather than go on which strategy they were taught to use, and Roberts, Gilmore and Wood (1994) indicate the difficulties in using introspection to reflect strategy use. External representations coupled with greater opportunities for strategic variety would provide more detailed data on the processes involved in reasoning.

Consistent with this difficulty in determining level of abstraction and expressivity is the paucity of opportunities for differences to be reflected. Measures in the reasoning tasks considered above have (with notable exceptions, such as Wood's (1969) questioning of the relationship between non-target individuals in the linear syllogisms task) focused on accuracy or latency measures. Errors or delays may be due to a whole panoply of processing differences. A task that can provide data from intermediate stages in the reasoning process would be instructive.

Up to now, only coarse variations in the appropriateness of different strategies to the task have been incorporated in reasoning studies. For example, Roberts, Wood and Gilmore (1997) compared strategy change from the one-person to the two-person compass directions task. The possibility of testing versions of the task that are intermediate in the appropriateness of the different strategies was not available. This is in part a result of the underspecification of the task requirements in terms of their computational properties, and so existing appeals to changes in memory load, or the amount of restructuring of the stimulus required are inadequate as these dimensions have not been fully determined. This reflects the primary requirement for a reasoning domain that emerged from the discussion of relating psychometrics to reasoning tasks: that preference for using different levels of abstraction be distinguished from ability to use abstract or concrete information to support problem solving. Developing and defining tasks where the problems can be altered so that different levels of abstraction may be equally appropriate for solving the problem would enable an assessment of preference for using different types of representation. In addition, changes in representation as the task requirements alter could then be assessed.

One such domain that fits many of these requirements is categorial syllogisms. The next chapter reviews individual differences in syllogistic reasoning, and indicates how it is a suitably rich task for exploring issues connecting strategies, representations and algorithms in reasoning. Up to now, the suggestion that different strategies reflect choices about the processes operating on representations with different levels of abstraction have been glossed over. This perspective on strategy variation is given an explicit treatment with regard to syllogistic reasoning. The remainder of this chapter discusses another suitable reasoning domain - Hyperproof. The discussion
will illustrate how the above requirements for suitable reasoning tasks are realised. In addition, the strategy distinction in terms of level of abstraction in representation will be highlighted by the domain, and new analyses of students' performance on Hyperproof will go some way towards detailing how cognitive load varies in different ways as problem requirements change and strategies vary.

3.3 Hyperproof and strategy

3.3.1 What is Hyperproof?

Hyperproof is a multimodal, computer-based logic course that aims to teach principles of analytical reasoning by combining graphical situations and sentential expressions of those situations in order to solve problems. Developed by Barwise and Etchemendy (1994) it was a product of their belief that reasoning ought to be heterogeneous, and furthermore more semantically- or situation-oriented. Larkin and Simon (1987) noted that inferences are often "effortless" from visual or spatial presentations of information, and Stenning and Oberlander (1995) have proposed that this is due to the inferences made during the construction of the diagram. Barwise and
Etchemendy intended to “exploit the power of visual representations in reasoning” (1990, p.18) by using both graphical situations and sentential propositional logic in order to solve reasoning problems.

Figure 3.5 shows the Hyperproof interface. The left window is the problem presentation. It consists of a graphical situation and below the graphics a list of statements about the situation in first order predicate logic (these will be termed “sentential expressions” in the following). The right column of this left window indicates that the statements shown are “given” to the student. Proofs are constructed in the sentential window using a Fitch-style proof system. Graphical situations can be incorporated into the proof by the use of operations that translate information from the sentential expressions to the graphical situation, or vice versa. The use of a graphical situation as part of the proof is indicated by a diamond icon as a step in the proof. Later in this chapter, when proof styles are discussed, a more detailed description of a proof will be given.

The graphical situation is given more expressive power by enabling the use of graphical abstraction. This can be done in four ways:

- **Position** can be left undetermined by placing the shape to the right of the chessboard. In Figure 3.5 the position of two objects is initially unspecified.
- **Shape** can be left undetermined by the use of a paper bag.
- **Size** can be left undetermined by the use of a cylinder. The cylinder can either be labelled with the shape of the object, or this too can be left undetermined. In Figure 3.5, two objects have both size and shape unspecified.
- **Items** have labels which can be left undetermined. In Figure 3.5, three of the six objects have no labels specified.

Combinations of these indeterminacies can be used to increase the abstractness of the situation.

On the right of the screen shown in Figure 3.5 are the goals of the problem. In the problem shown there are three separate goals that have to be solved. These goals can differ in terms of whether they require a graphical situation to be explored, or a sentential expression to be assessed. There

---

2 A fifth way is possible in HP, using a binary relationship “like/dislike” between shapes. However, this is not used in the problems solved by students in the studies reported here.
are two graphical goals and one sentential goal that have to be achieved in the problem shown. In order to suggest the appropriateness of different levels of abstraction for capturing features of the problem in Hyperproof, it is persuasive to imagine a full expression of the graphical situation shown in first order predicate calculus. It is equally compelling to think of representing the expression $\exists x \exists y (\text{Leftof}(x, y))$ in the graphical situation.

Studies of strategies and individual differences in reasoning have been conducted in HP. The following sections report the previous findings.

3.3.2 Learning and HP

Learning from HP was assessed by Stenning, Cox and Oberlander (1995) in a study that examined the transfer of reasoning skills from following the HP course compared to skills transferred from a more traditional syntactic course. This experiment is described in detail as the data from the experiment is reanalysed later in this chapter when strategic variation in using HP is discussed.

The aims of Stenning, Cox and Oberlander’s (1995) study were twofold: to provide some empirical observations of real logic teaching, and also to analyse the processes that determine transfer to other reasoning domains. 35 first year students from Stanford University participated in the study, attending a twelve week course on introductory logic. One group of students ($n = 22$) followed the HP course materials (Barwise & Etchemendy, 1994). Another group ($n = 13$) followed a more traditional logic course based on Bergman, Moor and Nelson’s (1990) text. This course used the HP interface for proof construction with the graphical window disabled.

Students were given two pre-tests to assess transfer to other reasoning tasks. One was the GRE test, as described above. The other test was a “blocks world” (BW) test which was based on the graphics of HP, but was expressed in standard English, so that no formal training was necessary in order to solve the problems. A graphical situation is presented along with certain constraints on the situation which are expressed in natural language. The BW test is described in greater detail in Chapter 7. Versions of these tests were also presented after the logic courses had been completed.
Learning from HP

Students were classified as GREA-Hi if they scored well on the GREA scale of the GRE pre-test, or GREA-Lo if they scored less well. For the BW test, GREA-Hi students that followed the HP course improved their scores on the post-test, but GREA-Lo students showed a decrease. For the non-graphical course, the opposite pattern occurred, with GREA-Hi students decreasing their score on the BW from pre- to post-test ($F(1, 26) = 9.45, p<0.01$), but GREA-Lo students increasing their scores.

For the GRE logical reasoning subscale (GREV), the non-graphically taught students improved significantly more than students that followed the HP course ($F(1, 28) = 4.93, p<0.05$). For the analytic reasoning subscale, students improved their scores no matter which course they followed, though there was a suggestion that the HP course increased scores more than the syntactic course.

The conclusions of this study showed that the HP course taught something different to students than did the non-graphical version of Hyperproof: the HP course provided transferable skills to analytic reasoning problems, the syntactic course provided skills that help in verbal/logical reasoning problems. Furthermore, the susceptibility of students’ learning from the course seemed to depend on their pre-test profile, i.e., whether they were GREA-Hi or GREA-Lo students. An analysis of the different use of HP made by GREA-Hi/GREA-Lo students provides a perspective on the way use of abstraction facilities varies for these groups of student.

Learning within HP

Observing the proofs that students produced in the exam at the end of the HP course, differences in the use of abstraction were apparent (Oberlander, Cox & Stenning, 1996). Figure 3.6 shows two alternative proofs that were constructed by students for one of the exam problems. The problem has four goals altogether, three of which are graphical, so about the shapes and sizes of the three objects that appear on the chessboard. The objects are represented with information only about their position. The fourth goal requires a sentential expression to be assessed. The diamond icons in the proofs indicate graphical situations that have been constructed by the student. A crucial feature of both proofs, and a necessary means for achieving the goals, is that several graphical situations are constructed in order to "exhaust" all the consistent extensions of the given graphical situation. This instantiates the "reasoning by cases" feature of HP (Barwise &
Etchemendy, 1994, pp.87ff.). An “exhaustive cases” structure can be seen in the proof in the left of Figure 3.6; this is indicated by a black bracket enclosing the diamond icons (each of which is a graphical situation). When all these possibilities have been expressed, the commonalities in all the situations can be expressed either graphically by the use of a “Merge” rule, or sententially by the use of “Inspect”. In Figure 3.6, each proof shows use of both “Merge” and “Inspect” following the “exhaustive cases” structure.

A distinctive feature of the two proofs illustrated is the different use made of these “exhaustive cases” structures. The proof on the left of Figure 3.6 is very “flat”, whereas that on the right nests exhaustive cases structures within a larger exhaustive cases structure. The left proof makes use of 10 graphical situations (marked by “Assume”), whereas the proof on the right makes use of 8 assumed graphical situations. Furthermore, the graphical situations constructed in the left-hand proof are more concrete, in that for each situation the size and shape of each object is specified. The right-hand proof employs graphical situations that leave several features of the shapes unspecified. This is a concomitant property of the nesting structure, where features of the objects are specified gradually, with situations in more deeply nested structures defining features of the blocks that are previously left unspecified. These two proofs are examples of two general proof types: those that use graphically abstract situations and nesting, and those where situations
are more concrete and little or no nesting occurs.

These structural differences have been shown to have correlates in terms of order of rule use where certain bigrams and trigrams of rule use are indicative of one or other of the proof styles (Monaghan, 1995). These differences were emphasised when use of graphical situations were distinguished into those that specify all the graphical information ("full-assume") and those that leave some graphical information abstract ("part-assume"). The relative frequency of use of these rules both as unigrams, bigrams and trigrams are hallmarks of the two proof styles. The amount of nesting in the proofs, the mean concreteness of the situations depicted, and the ratio of full-assume to part-assume use are all strongly related (Monaghan, 1998b), so the general features of the proof that are distinguished in multiple ways can be reflected in a single measure.

The proof style differences shown in Figure 3.6 are indicative of the proofs produced by GREA-Lo and by GREA-Hi students. The left-hand proof is by a GREA-Lo student, the right-hand proof is by a GREA-Hi student. GREA-Hi students seem to use more nesting in their proofs than GREA-Lo students, though this was not found to be significant (Oberlander, Cox & Stenning, 1996). GREA-Hi students use more part-assume rules: 44% compared to GREA-Lo's 36% (Monaghan, 1995), and their graphical situations contain more graphical abstraction ($t(18) = 1.91, p<0.05$) than those used by GREA-Lo students.

The link between the aptitude-treatment interaction in learning from the different logic courses, and the difference in the strategies used by GREA-Hi and GREA-Lo students led to the formulation of an "abstraction ability hypothesis" (Oberlander, Cox, Monaghan, Stenning & Tobin, 1996). This contends that GREA-Hi students benefit from the HP course because they are better able to manipulate the graphical abstractions it offers. GREA-Lo students do not use graphical abstractions to the same extent, and so the level of abstraction in their proofs may not be so optimal for solving the given problem. This hypothesis gains support from the proof style analyses reported above and by the observation that GREA-Hi students operate over graphical situations when reasoning, whereas GREA-Lo students just output graphical situations. Hence, the GREA test is a reflection of different uses of levels of abstraction in representation.

However, as yet the "appropriateness" of the level of abstraction in the representations used to solve the problem has not been addressed adequately. This is because no analysis of the task requirements of HP problems has been undertaken. In the above studies it was assumed that more abstraction in the task is better for solving reasoning problems found in the GREA or the
HP course, but the previous chapter indicated that developing skills for solving reasoning problems requires finding the right level of abstraction in the representation, and that this is a balance between the effort required to unify and store the information in the problem and the effort required to operate on more abstract information and reprocess information when necessary. The following section analyses two questions from the HP exam that provide a quantifiable assessment of the level of abstraction used and the changing memory load requirements for different strategies used by students. The GREA-Hi/GREA-Lo distinction is shown to be a measure of a dynamic approach to problems rather than a propensity to use graphical abstraction mechanisms in reasoning.

3.4 Strategy change in HP

Students following the HP course in the study described above sat an exam that contained two problems with different degrees of indeterminacy in the graphical situation. Of the other three problems on the exam, two contained little or no graphical abstraction and the other was at the end of the exam and few students managed to attempt a solution. No strategic differences in proof style were found for the non-abstract problems, so they are omitted from the following analysis.

The two graphically abstract problems differed in the degree of abstraction that the student has to work with in order to achieve the goals. One was the problem shown in Figure 3.6. This problem (call it high-abstract) requires the student to consider the size and shape for all three blocks in order to attain the goals. If no graphical abstraction is used, it requires 10 situations to be constructed in order to achieve the goals. If graphical abstraction is used in the situations then fewer situations are required. The second problem is shown in Figure 3.5 and though it contains the same degree of graphical abstraction (two blocks with size/shape unspecified, and two blocks with position unspecified) this problem requires fewer of these variables to be considered for achieving the goals. The goals in this question only require the position of the two blocks and label of one shape to be considered, and this can be done in three situations. It is therefore referred to as a “low-abstract” problem.

For the high-abstract problem, as previously mentioned, the use of graphical abstraction reduces the number of situations required, as then the cases relevant to achieving the goals can be covered
at a higher level of "granularity". In addition, a nesting structure means that the information presented in the sentential expressions can be applied systematically, and this systematicity is externalised. For the GREA-Lo students, the application of information from the sentential statements has to be internalised, and so memory load is greater. The nested structure proofs reduce the number of operations that have to be made during situation construction, but may require extra effort in structuring the proof prior to construction. Thus, the level of abstraction present in representations, and the level of abstraction required in order to solve the problems can be expressed exactly in HP, and can be tied to different processing requirements. Also, HP indicates the different actual use of abstraction by students in their situation construction, and this satisfies the several requirements on a reasoning domain to provide a quantifiable description of level of abstraction used in different strategies, one consequence of which is to distinguish issues of style and ability in strategy use.

For the low-abstract problem, in contrast, only two labels have to be considered, so memory load is very low and any external structuring of the proof is not offset by a significant reduction in load. Thus, an appropriate level of abstraction in the representations for the high-abstract problem will be higher than for the low-abstract problem. If the GREA measure relates to appropriate use of graphical abstraction then different strategies for the two problems will be seen for the GREA-Hi students, and different patterns of inappropriate strategies will be seen for GREA-Lo students. In the following analyses, GREA score is treated as a dependent variable, and strategy choice on each question as factors. A version of the first part of the following results was presented in Monaghan (1998a; 1998b).

The two proof types depicted in Figure 3.6 can be described in terms of the serialist/holist framework. A proof was classified as being "serialist" if more than half of the graphical situations were full-assumes, and being "holist" if half or more than half of the graphical situations were part-assumes. These terms are used to reflect the relationship between step-by-step situation construction strategies and the serialist learning style, and between the structured approach which uses more abstraction and the holist learning style. The link is theoretically motivated as no tests of serialist/holist learning style were administered.

The flexibility of students in strategy employment related to scores on the GREA pre-test. Students who were flexible in their strategy use scored higher on the pre-test, whereas those that

---

3In the previous analyses of this HP data GREA groups were entered as a factor.
did not vary their proof strategy scored lower (F(1,20) = 6.42, p<0.05). Not varying strategy when that strategy becomes inappropriate reflects the "pathological" learners that Pask (1976) observed, and also relates to the inflexible use of strategies by the subjects in Roberts et al.'s (1997) study of the compass directions task.

<table>
<thead>
<tr>
<th>Strategies used</th>
<th>N</th>
<th>Mean GREA score</th>
</tr>
</thead>
<tbody>
<tr>
<td>low-abstract</td>
<td>high-abstract</td>
<td></td>
</tr>
<tr>
<td>serialist serialist</td>
<td>11</td>
<td>7.27</td>
</tr>
<tr>
<td>holist holist</td>
<td>4</td>
<td>4.00</td>
</tr>
<tr>
<td>serialist holist</td>
<td>5</td>
<td>9.60</td>
</tr>
<tr>
<td>holist serialist</td>
<td>2</td>
<td>5.50</td>
</tr>
</tbody>
</table>

Table 3.1: Strategy change and GREA score.

A simple four-way ANOVA proved significant for GREA score by strategy (F(3, 18) = 5.63, p<0.01). A Scheffé post hoc test indicated that the effect was due to the difference in means between holist-holist and serialist-holist students (see Table 3.1). This effect must be treated with caution due to the small number of students in the holist-serialist group. Students that seem to perform best on the GREA test use a serialist strategy on the low-abstract problem and a holist strategy on the high-abstract problem. Thus, the GREA scale does not reflect a particular approach to using graphical abstraction, rather it reflects appropriate use of graphical abstraction: GREA-Hi students seem to be able to select an appropriate level of abstraction in the representations they use. Presumably this skill of selection means that the problems with varying levels of abstraction in the GREA test can be solved appropriately, leading to higher scores on this test.

Students that use the serialist strategy on both problems in HP score high on the GREA, these students seem to be limited to using only one strategy type – where graphical situations tend to be concrete – but these students use it effectively in solving problems even when extra effort is required. It is an unexplored issue as to what would happen if the amount of information that had to be considered was even greater, if the number of situations far exceeded working memory storage capacity, for example.

The strategy groupings can also be used to analyse change in test scores from pre- to post-test.
on the GRE and the BW tests. Repeated measures ANOVAs were performed, with score on the test as dependent variable, and time (pre/post) as a within-subjects factor. Strategy types on low-abstract and high-abstract problems were entered as between-subjects factors.

<table>
<thead>
<tr>
<th>Strategies used</th>
<th>N</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>low-abstract</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>serialist</td>
<td>7</td>
<td>5.00</td>
<td>5.86</td>
<td>0.86</td>
</tr>
<tr>
<td>holist</td>
<td>3</td>
<td>4.67</td>
<td>4.00</td>
<td>-0.67</td>
</tr>
<tr>
<td>serialist</td>
<td>4</td>
<td>6.25</td>
<td>4.75</td>
<td>-1.75</td>
</tr>
<tr>
<td>holist</td>
<td>2</td>
<td>6.00</td>
<td>3.00</td>
<td>-3.00</td>
</tr>
</tbody>
</table>

Table 3.2: Strategy change and GREV score.

For the GREV test, there were no main effects of strategy type. Time approached significance (F(1, 12) = 3.38, p=0.09), and an interaction between the two strategy types and time also approached significance (F(1, 12) = 3.96, p=0.07). Means for pre- and post-tests are shown in Table 3.2. Only students that use little graphical abstraction (serialist-serialist strategy) seemed to improve their scores on the GREV test (t(14) = 2.29, p<0.05).

<table>
<thead>
<tr>
<th>Strategies used</th>
<th>N</th>
<th>Pre-test</th>
<th>Post-test</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>low-abstract</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>serialist</td>
<td>7</td>
<td>7.43</td>
<td>8.86</td>
<td>1.43</td>
</tr>
<tr>
<td>holist</td>
<td>3</td>
<td>3.33</td>
<td>6.67</td>
<td>3.33</td>
</tr>
<tr>
<td>serialist</td>
<td>4</td>
<td>9.50</td>
<td>9.75</td>
<td>0.25</td>
</tr>
<tr>
<td>holist</td>
<td>2</td>
<td>3.50</td>
<td>2.50</td>
<td>-1.00</td>
</tr>
</tbody>
</table>

Table 3.3: Strategy change and GREA score.

For the GREA test, there is a main effect of strategy on the low-abstract problem (F(1, 12) = 12.29, p<0.005). This is a reflection of the low GREA score of students that used an inappropriate holist strategy on the low-abstract problem, a tendency that is reflected in low scores on the post-test also. No interactions with time were significant, but increases in score were variable among the strategy choice groups (see Table 3.3).

4Pre- and post-test data was only available for 16 of the 22 students.
<table>
<thead>
<tr>
<th>Strategies used</th>
<th>BW</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
</tr>
<tr>
<td>serialist</td>
<td>7</td>
</tr>
<tr>
<td>holist</td>
<td>3</td>
</tr>
<tr>
<td>serialist</td>
<td>4</td>
</tr>
<tr>
<td>holist</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.4: Strategy change and BW score.

For the BW test, there was a main effect of strategy choice on the low-abstract problem (F(1, 12) = 8.40, p<0.02), but no interactions were found with time (see Table 3.4).

These results show that when variations in strategy in terms of the level of abstraction used in the representations is investigated, more subtle changes than were initially proposed can be perceived in the transfer of reasoning skills from HP. Students that use a serialist, or concrete, strategy for both low- and high-abstract HP problems are the only group to find HP training beneficial for reasoning on both the GREA and the GREV subscale. The students that used a serialist/concrete strategy for the low-abstract and a holist/abstract strategy for the high-abstract problem were predicted to benefit most from HP. This was not the case, as their score seemed to decrease for the GREV subscale. This means that these students learned a particular skill which proved detrimental to solving certain reasoning problems, problems where specifying information is not so helpful\(^5\). HP taught these students different ways of approaching problems with varying numbers of cases, but it did not seem to teach useful approaches to problems when indeterminacy is very high.

The students that use the holist/abstract strategy on the low-abstract problem and the serialist/concrete strategy on the high-abstract problem have learned a flexible approach from HP also, but they have not learned appropriate application of strategy to the requirements of the problem. These students get worse on both the GREV and the GREA subscales. Generalising from the behaviour of only two students is tendentious, however, so any conclusions regarding these students' performance are only suggestive.

Finally, the students that use the holist/abstract strategy on both low- and high-abstract problems

\(^5\) and may even be detrimental (see Chapter 7).
seem to become very good at solving one type of problem: their scores on the GREA increase dramatically from being the lowest, to being closer in line with the better reasoners. However, this improvement seems to be at the expense of their ability to solve GREV problems, where their score deteriorates, as it also does on the BW test. These students seem to have benefitted from HP's graphical abstraction facility, which they can use well on one type of problem, but which interferes with their ability to solve other reasoning problems where graphical abstraction is less appropriate.

HP teaches different skills to different students. Students that remain impervious to the graphical abstraction mechanisms of HP, paradoxically, seem to benefit in a general way from the course. Students who learn to apply graphical abstraction appropriately also benefit on the BW and the GREA test, but their performance is impaired on the GREV problems where graphical abstraction cannot be used. Some students seem to learn to use graphical abstraction to overcome problems they initially had with determinate problems and so their score increases on the GREA scale, but the price of this is a reduction in their ability to solve GREV problems.

The GRE test proves to be valuable as a gauge of strategy use with regard to the level of abstraction employed in graphical situations. It is also useful as a reflection of the effects of learning to use graphical abstraction facilities. The GRE in close comparison with HP illustrates how understanding of the domain assists in a deeper understanding of the measure.

The analysis of HP performance indicates that the suggestions of the role of different levels of abstraction in representation observed in different strategies for various reasoning tasks is a productive approach to classifying the use and development of strategies. Furthermore, it has been shown that this can be done in a quantitative way, which shows that different use of representations relates to different strategies, and that these strategies can be classified in terms of being serialist or holist. However, this initial stab at relating psychometric approaches to computational accounts of individual differences in reasoning requires further elaboration. In particular, relating HP performance to psychometrics directly would be useful, providing empirical support for the aligning of serialist/holist styles with GREA scores, and, in addition, deriving further data for what spatial ability measures are testing. In addition, a more detailed account of exactly what the GRE test is actually probing is necessary if a full account of style is to be provided. Also, further justification and exploration of the processes required by representations using different levels of abstraction is necessary, tying in the strategy - representation accounts with descriptions
of the algorithms underlying the representation selection.

One major shortcoming of the HP domain for studying reasoning is that the internal reasoning of the student is not directly reflected, only the external representation use is recorded. The importance of this becomes apparent when the inflexible, serialist strategy students are shown to have learned the most from the HP course. These are students that "output" graphics rather than work with the graphical representations, so the degree of their internal structuring is not apparent. Two ways of getting round this problem are to use psychometrics to show that patterns of response are generalisable. Another method is to stipulate the external representation to be used by the student and then assess the reaction of the student to the enforced strategy. This approach is taken when methods for teaching syllogisms are given to students. The relation between this teaching study and the free-choice environment of HP will indicate the extent to which the HP results are generalisable to other reasoning domains, and the relationship between internal and external representation in HP and other reasoning domains.

Chapter 5 discusses the approach to be taken in order to extend this initial HP study, and presents the hypotheses to be tested. Chapter 4 discusses categorial syllogisms as a potentially useful domain for detailing the processes and preferences dependent upon different representation use. This next chapter will also discuss why categorial syllogisms are good for testing style of representation selection in addition to measuring ability to use abstract representation.
Chapter 4

Representations, strategies and algorithms in syllogistic reasoning

4.1 Introduction

Categorial syllogisms offer a useful domain for studying the relationship between strategies, representations and algorithms in that, after (or despite) over 2000 years of study, the task is relatively well-understood. In addition, they are a fruitful area for individual differences in response. For some problems almost no subjects get the problems right, for others almost every subject gets the correct answer (Johnson-Laird, 1983). Differences in response are also qualitative, in that different strategies that seem to invoke different representations are spontaneously developed by students solving the problems. Stenning and Oberlander (1995) state that there is "a quite general weakness of psychological theories of reasoning which generally focus on the processes which follow representation selection" (p.32), and this weakness is a failure to analyse how and why particular representations are selected or recognised as applicable by the subject. Of particular importance is that two diverse strategies develop as a result of exposure to the task. In this respect it is distinct from the range of strategies discussed in Chapter 2 for other reasoning domains, where use of different strategies have a "privilege of occurrence". This alone suggests that the types of strategy used in syllogism solution are equally appropriate for the task, though the different strategies are better or worse depending on the exact problem. The range of strate-
gies demonstrated in syllogism problem solving is therefore appropriate for investigating style of reasoning as well as ability.

Syllogisms are reasoning problems with two quantified statements, the first statement links properties A and B, the second links properties B and C. The subject's task is to say what, if anything, follows from the information given, i.e., what connection can be formed between properties A and C. Statements can take one of four forms:

- All As are Bs
- No As are Bs
- Some As are Bs
- Some As are not Bs

and the properties can be in any order. In the standard task, the subject is asked to assume that there exists at least one A, B and C. There are 64 different premiss pairs, 27 of which render conclusions which can be expressed in the form of one of the four statements.

Original presentations of the problems presented either multiple-choice answers or lists of statements that the subject had to check. Johnson-Laird and Steedman (1978) presented the problems in a novel manner, where the subject was required to "say what follows", without any requirement to make a multiple choice response. This change from assessing answers to providing answers provided a greater variety of responses. They found that for some syllogisms, such as "Some As are Bs, All Bs are Cs" almost all subjects got the right answer ("Some As are Cs"). For other syllogisms, such as "All Bs are As, No Bs are Cs" virtually no subjects got the right answer ("Some Cs are not As"). A variety of accounts have been proposed to account for these variations in response.

Conventionally, there are three different accounts of representations during syllogistic solution. One proposes that internal representations are spatial (Erickson, 1974). Erickson suggests that subjects use Euler’s Circles to represent the relations between properties in the syllogistic premisses. As there are several arrangements of circles that make each premiss true, subjects consider only a limited number of these arrangements.

A second theory holds that subjects reason by virtue of mental rules, similar to steps in a deductive proof. No fully explicit account has been postulated, but Braine and Rumain (1983) and
Rips (1994) have argued for inference rules in a range of problem solving domains which can be extended to syllogisms.

The third contrasting account states that subjects construct models of individuals having certain properties in accordance with the premisses (Johnson-Laird & Steedman, 1978; Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991).

Subjects construct a set of models of the premisses that makes explicit the minimum amount of information; they formulate a parsimonious conclusion based on this set; and to test for validity, they search for counterexamples, perhaps fleshing out the initial models in order to do so (Johnson-Laird & Byrne, 1991, pp.118-9).

There are several other accounts of syllogistic reasoning which posit that subjects are influenced only, or primarily, by the surface features of the premisses, rather than that they attempt to reason about the relations of the premisses. Sells (1936) suggested that subjects respond to syllogistic problems according to the “atmosphere” of the premisses. Thus, if one or more of the premisses are negative then the conclusion will be negative. If one or more of the premisses are particular then the conclusion will be particular. Wetherick (1989) has suggested an even simpler “surface feature” strategy for solving syllogisms which, unlike the atmosphere hypothesis, does not involve combining information from both premisses. Wetherick proposes that subjects make a conclusion in line with the more conservative of the premisses, where the order of conservatism is (from most to least): no, some...not, some, all.

However, all these accounts of responses to syllogisms are qualitative accounts – variation is along one proposed dimension. That subjects demonstrate very different strategies as they are solving syllogisms provides an argument against such theories of syllogistic reasoning, showing that a single-framework theory at best “accounts for how some people reason some of the time” (Johnson-Laird & Byrne, 1991, p.139).

4.2 Strategies in syllogistic reasoning

There are two principle studies of syllogistic reasoning that have recognised and explored different strategies for solving the problems. The first was Galotti, Baron and Sabini’s (1986) investigation into use of “short-cut” rules for solving syllogisms, which follows on from discussions of strategy change in Chapter 2. The second is Ford’s (1995) investigation of subjects that solved
syllogisms using either a "spatial" or a "substitution" strategy, where use of different strategies is best described as a bifurcation of methods as with proof style in HP, rather than a linear progression as in the studies discussed previously. These strategies bear strong resemblances to the two methods for solving syllogisms compared by Stenning and Yule (1997). In addition this study provides the strongest evidence for the vulnerability of syllogistic reasoning to strategic variation, describing strategies which seem to rely on very different forms of representation. This evidence suggests the suitability of syllogisms for investigating the relationship between representation and strategy and deriving a computational account of the different processes to which these representations give rise.

Galotti, Baron and Sabini (1986) asked their subjects to report on the procedures that they used in order to solve syllogisms. Evans, Barston and Pollard (1983) also used protocols of subjects' introspections, but their problems were presented with conclusions that had to be assessed, inducing different processes to those in Galotti et al.'s experiment where subjects had to generate their own conclusion. They distinguished three groups: poor reasoners who were selected from a general undergraduate population and scored low on a pre-test of syllogisms; good reasoners, who were selected from the same population but who performed well on the pre-test syllogisms; and expert reasoners, who were graduate students with some training in logic. Each subject was given all 16 premisses of the first mood (so premisses ordered A-B, B-C). In their first study, good and poor reasoners were compared. Good reasoners were more accurate and faster than poor reasoners. Comparing discussions on how each subject solved the problems revealed that good reasoners were more likely to discuss the use of "short-cut" rules. For example, subject MK reports:

...and I realized that, when there's a *some* and a *some*, nothing ever follows (Galotti *et al.*, 1986, p.19).

In the standard task, syllogisms with two particular premisses ("two somes") or with two negative premisses ("two negatives") have no valid conclusion. Galotti *et al.* suggest that the good reasoners' better performance on the problems was due to the use of these short-cut deduction rules. Most of the rules stated by subjects were formulations of the "two-somes" rule but some instances of "two-negatives" rules occurred. Dividing the syllogisms with no valid conclusion into those where the two-somes rule applies, those where the two-negatives rule applies, and
other problems did not reveal a difference between good and poor reasoners. However, in a second study when expert reasoners were also included, experts were significantly better than good and poor reasoners on the "two-somes", "two-negatives" and "other" problems, and good reasoners were significantly better on the "two-somes" and the "two-negatives" problems than poor reasoners.

Use of the short-cut rules in syllogistic reasoning fits the observations of Roberts (in press) for the general nature of strategic change in a range of reasoning tasks. Expert reasoners are those that use short-cut rules when appropriate. But they are also better at problems where these rules do not apply. Their ability is selective use of rules when appropriate. As these rules operate on the grammatical form of the premisses they can be seen to be verbal, but this classification does not serve as a distinction with other methods: the approach of novices in Galotti et al.'s task is under-specified. Besides fitting the generalisations of experts' patterns of reasoning in a number of domains, this study also points to the variety and importance of strategy in syllogistic reasoning. Distinguishing problems according to the strategies that operate upon them is instructive in the case of syllogisms where the two-somes and the two-negatives rules apply. Different subjects respond differently to these problem types. Ford's (1995) study extends the investigation on strategic development begun in Galotti et al.'s study, where the "verbal" strategies can be seen to be embedded in a more general strategic framework.

Ford presented all 27 valid syllogisms to her subjects in individual sessions, and asked them to speak aloud as they solved the problems. After completing all the syllogisms, the subjects went over all the problems again, explaining to the experimenter the method they used to solve the problems. During the course of solving the problems, most subjects began to develop spontaneously one of two distinct strategies as they solved the problems. Ford terms these "spatial" and "substitution" strategies. Of the twenty subjects in the experiment, eight demonstrated "extensive use" of spatial representations but did not use, or rarely used, substitution. Eight subjects used substitution, and never used spatial representations. Two subjects used mixed strategies, and two subjects struggled with the task so could not be classified as using one or other strategy. The criterion for being classified as using a spatial strategy was "when the subject drew shapes in different spatial relationships to represent the relationships in the syllogisms" (Ford, 1995, p.18). The criterion for using a substitution strategy was when one of the following occurred: "a subject spoke of replacing one term with another, a term in a syllogism was crossed out and replaced
with another, arrows were drawn between terms in the syllogism, the syllogism was rewritten as an equation or as terms connected by arrows" (ibid.). Ford (1995) refers to the substitution strategy as "verbal" (p.19) and this will be the term used to describe this strategy from now on.

The two strategies are now described, and the essence of the empirical results used to support the theoretical distinction are summarised. The strategies are presented in detail, as their relationship to the theoretical methods for solving syllogisms presented later in the chapter requires this depth of analysis.

The verbal strategy is characterised by some form of substitution, where a term from one premiss is implanted in the other premiss. One subject uses an algebra metaphor to describe the procedure: "we can just like plug it right in it's like doing algebra" (Ford, 1995, p.17). Ford proposes that this verbal manipulation of the premisses takes the form of one premiss providing the value for substitution (premiss relating class C to a property P) while the other premiss provides the term that needs to be substituted (containing specific objects O whose status with regard to either C or P is known). The principle that subjects are struggling to acquire is expressed below:

A. If a rule exists affirming of every member of the class C the property P then
   (i) whenever a specific object, O, that is a member of C is encountered it can be inferred that O has the property P
   and
   (ii) whenever a specific object, O, that lacks property P is encountered it can be inferred that O is not a member of C
B. If a rule exists denying of every member of the class C the property P then
   (i) whenever a specific object, O, that is a member of C is encountered it can be inferred that O does not have the property P
   and
   (ii) whenever a specific object, O, that possesses the property P is encountered it can be inferred that O is not a member of C.

A.(i) and B.(i) use modus ponens whereas A.(ii) and B.(ii) use modus tollens. It will be seen later that this method closely resembles the ND method developed by Stenning and Yule (1997).

¹In actual fact, this is not exactly modus tollens, as it permits A → ¬ B and B to give ¬A, rather than limiting to the A → B and ¬B to give ¬A case. In the following, by referring to modus tollens I mean this more abstract version of the rule.
Syllogisms can either be solved using simple substitution (when the value to be substituted is readily available) or using sophisticated substitution, when the premiss containing the value to be substituted requires some manipulation. Syllogisms that utilise the same rules given above as well as the same type of substitution (simple or sophisticated) tend to be solved to the same degree of accuracy by the group using the verbal strategy. Also of great importance is the order of terms in the premisses. For example, the syllogism “Some Bs are As, All Bs are Cs” is harder for the verbal strategy subjects than “Some As are Bs, All Bs are Cs” (46.6% compared to 93.3%).

Classifying syllogisms along these lines does not make sense for the spatial group (their strategy is described below) as the accuracy of their response to syllogisms that require similar substitution operations can vary radically, and the order of terms in the premisses, when it does not affect the number of models consistent with the premisses, does not influence accuracy in the spatial group (accuracy for “Some Bs are As, All Bs are Cs” is 68.7% whereas that for “Some As are Bs, All Bs are Cs” is 87.5%, a much smaller difference than that for the verbal subjects).

The spatial method is characterised by subjects’ development and use of graphics to indicate different classes and the relationship between these classes. These bear a close relationship to Euler’s Circles (Ford, 1995, p.41). The difficulty of syllogisms to the spatial group was determined by the number of possible arrangements of circles resulting from the integrating of the information in the premisses. For example, syllogisms where the diagrams are fully constrained are the easiest (e.g., “All As are Bs, No Bs are Cs” – the A is contained in the B circle and the C circle lies outside the B circle), and responded to homogeneously by the spatial group (for these 6 syllogisms, the range was 80.0% to 100% accuracy). The verbal group responded with varying accuracy to this set of syllogisms (60% to 100%).

In using graphics to represent the premisses, there is a difficulty in combinatorial explosion in the arrangements of circles to represent the premisses, a valid criticism directed towards theories of Euler’s Circles as internal representations during syllogism solution (Johnson-Laird & Byrne, 1991). However, in order to use graphics effectively to support their reasoning, several subjects attempted to use an abstraction mechanism so that one diagram represents several models – this is akin to use of the cross-notation in order to represent the minimal model in the Euler’s Circles method described below.

The distinction between these two methods has been described by Ford in terms of representational differences. Another way of describing the differences between these two strategies is in
terms of the verbal strategy being a *syntactic* approach and the spatial strategy being *semantic*. The verbal strategy is syntactic because it operates on the grammar of the premisses, whereas the spatial method is semantic because difficulty in using it is determined by the number of models consistent with the premisses. Stenning (1999) provides a sceptical response to this characterisation, indicating that a mechanism for a semantic system is indistinct from that for a syntactic system for reasoning.

An alternative description of the differences is in terms of the strategies, and this begins to abstract away from the representations themselves. The representations induce different processes, but the focus is on the process rather than the representation. These different processes will be seen to be those required by the algorithm underlying representation use. Clarifying the nature of this relationship is one aim of this chapter, and to this end two theoretically motivated methods for solving syllogisms are discussed and related to Ford’s strategies. This accumulates support for the contention that differences between Ford’s strategies can be profitably considered in computational as well as representational terms. A thorough description of the two methods is also important as they are used in Chapter 6 in the teaching syllogisms experiment, where responses to the two methods are compared to the psychometrics considered in Chapter 3.

4.3 Two methods for solving syllogisms

Debates over what form the fundamental reasoning mechanism takes have used the syllogistic task as its battlefield. The spatial, rule-based, and mental models accounts have all been proposed within this context. One major misconception in the debate has been over exactly what are the computational differences between seemingly conflicting accounts (see Chapter 1). Stenning and Oberlander (1995) developed an Euler’s Circles (EC) method for reasoning with syllogisms to illustrate that mental models is only one implementation of an abstract algorithm: at the “logic” level of description, the EC method and mental models are equivalent, they differ in terms of implementation. Current approaches to studying syllogisms cannot distinguish between competing accounts, as data consisting of input-output pairings cannot decide between one implementation over another. It is a moot point as to whether intervening in the process of syllogistic reasoning can provide any discriminating data to settle the debate (see Bucciarelli & Johnson-Laird, in press). Their study shows that subjects can construct counterexamples to given conclusions that
do not follow from premisses, and this counterexample search is an essential part of the mental models reasoning process. However, the finding that subjects can construct counterexamples is very different to them actually doing so during syllogism solution, the requirements are different in the two tasks and may induce different strategic approaches. More critically, these subprocedures of mental models reasoning could equally well be implemented in a number of different ways. A method using Euler’s Circles could be adapted to involve construction of counterexamples. The commonality between the EC and the mental models method was further explored by Stenning and Yule’s (1997) implementation of the same algorithm in an adapted fragment of propositional calculus, with a small set of natural deduction rules. Hereafter, this is referred to as the Natural Deduction (ND) method. As with the EC method, the ND method can also be formulated so that counterexamples are constructed. Stenning and Yule’s aim in comparing these different implementations of the same algorithm was to indicate that a class of fragments of systems are equivalent. However, differences do arise between individuals, and these differences were proposed to be in terms of the larger systems the fragments are drawn from. So, the system from which the EC method is drawn from will induce different processes, or ways of thinking from the system that the ND method is drawn from. Below, the EC and the ND methods are shown to be slightly different algorithms, but differences that are the result of being fragments of the two larger systems.

Both the EC and the ND method are implementations of the “identify critical individuals algorithm” (ICIA). The details of this algorithm will be discussed with respect to the detailed descriptions of the EC and the ND methods below. Those with a spatial preference are referred to the EC method, for those with a verbal preference the ND method discussion ought to be more illuminating, whilst for the flexible reader either but preferably both will serve.

This underlying algorithm provides the computational similarity between the two methods: in Stenning and Oberlander’s (1995) terms, the logic is the same, it is only the implementation that differs. For every operation in one method there is a corresponding operation in one-to-one correspondence in the other method. This means that differences in response to the two methods are due to differences in implementation, rather than differences in the task. As with other reasoning domains discussed in the last two chapters, the focus here is on strategic variation revealed by use of different representations.
4.3.1 Euler's Circles method

In this method, Euler’s circles are used to represent the information given in the premisses of the syllogism. Each property is represented by a circle, and each premiss is represented as shown in Figure 4.1. For each premiss the characteristic diagram is used which represents the maximum number of types of individual consistent with the premiss. Stenning and Oberlander (1995) call this the **maximal model** for the premiss. This representation avoids the combinatorial explosion of diagrams that would have to be processed if all Gergonne relations were used for each premiss (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991). Areas of the diagram that are known to be non-empty represent **minimal models** of the premisses, and are indicated by a cross in the diagrams. Hence, the Euler’s Circles representations indicate an abstraction of one diagram across several models.

![Figure 4.1: EC representations of syllogism premisses.](image)

The two diagrams resulting from translating the premisses then have to be integrated to produce the maximal model for the three properties. The two B circles are made to correspond, then the A and the C circles are arranged so as to maximise the number of regions which are consistent with the premisses. If a minimal region is divided by the third circle, then the cross is removed.
The remaining crosses indicate the critical regions, corresponding to a maximal type of individual which must exist. As the EC method hinges on identifying these critical individuals it is seen to be an implementation of the ICIA.

All syllogisms which have valid conclusions have integrated diagrams with maximal types. Forming a conclusion requires describing the critical individual. Universal conclusions follow if the critical individual is in an unbroken circle which represents an end term, in which case the unbroken circle becomes the subject term of the conclusion. If there is no such region, there is no valid universal conclusion.

Figure 4.2 indicates the EC method being used to solve the syllogism “No As are Bs, Some Bs are Cs”. Note that when the A and the C circles are arranged, they intersect, so that the cross from the A circle is removed. There is just one cross remaining, so this translates as a critical individual with properties ¬ABC, and the conclusion is existential as the critical region is not circular.

\[
\begin{align*}
\text{No As are Bs} & \quad \begin{array}{c}
A \quad \text{x} \quad B \quad \text{x}
\end{array} \\
\text{Some Bs are Cs} & \quad \begin{array}{c}
B \quad \text{j} \quad C
\end{array} \\
\text{A} \quad \text{x} \quad \begin{array}{c}
B \quad \text{z} \quad C
\end{array} \\
\text{conclusion:} & \quad \text{Some Cs are not As}
\end{align*}
\]

Figure 4.2: EC method used to solve the syllogism “No As are Bs, Some Bs are Cs”.

4.3.2 Natural Deduction method

The syllogistic system can be expressed in monadic predicate form (see Johnson-Laird & Byrne, 1991, p.117). Omitting quantifiers provides a further simplification which more clearly indicates the propositional operations. For further details, see Yule (1995). The premisses are represented
The ND rules required to solve all syllogisms are as follows:

- (1) **modus ponens** If we have \( A \rightarrow B \), and we have \( A \), then we can get \( B \).

- (2) **modus tollens** If we have \( A \rightarrow B \), and we have \( \neg B \), then we can get \( \neg A \).

- (3) **conjunction elimination** If we have \( A \& B \), then we can get \( A \), and we can get \( B \).

- (4) **conjunction introduction** If we have \( A \), and we have \( B \), we can get \( A \& B \).

- (5) **double negation elimination** If we have \( \neg\neg A \), we can get \( A \).

Applying rules according to a given procedure means that the method becomes a decision procedure for solving syllogisms. Such a procedure is given in Table 4.2 (adapted from Stenning & Yule, 1997). In the teaching syllogisms study, a method very similar to this was used (though terms like "existential", "end-term", "antecedent" were replaced with descriptions that did not presuppose logical training).

The method was given to the subjects for reference as they worked on the syllogisms. Essentially, it involves 4 stages:

- find the source premiss: this refers to the premiss that the initial description of the individual is made from. A decision has to be made about which premiss to begin with;
1. Seek a unique existential premiss:
   (a) If there are two, then respond NVC and quit.
   (b) If there are none, then go to 2.
   (c) If there is a unique one, make it the source premiss and go to 4

2. Seek a unique positive universal-premiss end-term subject:
   (a) If there are none, then go to 3.
   (b) If there is one choose its premiss as the source premiss and go to 4.
   (c) If there are two, conclude NVC.

3. Seek a unique positive universal-premiss middle-term subject:
   (a) If there are none, then conclude NVC.
   (b) If there is one choose its premiss as the source premiss and go to 4.
   (c) If there are two, choose an arbitrary source premiss.

4. If the source premiss is existential, then take its two terms as the first two clauses of the individual description. If the source premiss is universal, assume its antecedent. Apply modus ponens and conjoin the consequent to the antecedent to make the first two clauses of the individual description.

5. Compare middle terms:
   (a) If a source middle term matches (with regard to negation) the antecedent middle term of the conditional premiss, apply modus ponens, and conjoin consequent term to individual description. Go to 6.
   (b) If the source middle term mismatches (with regard to negation) with the conditional consequent middle term, apply modus tollens to the conditional premiss, and conjoin consequent term to individual description. Go to 6.
   (c) ELSE conclude NVC and quit.

6. Individual description is now complete.

7. Draw Abstract conclusion from individual description:
   (a) Delete B conjunct from individual description. Quantify existentially for an existential conclusion (reordering any positive conjunct into subject position).
   (b) If clause 2b was satisfied, then there is a universal conclusion with the source premiss end-term as subject.

Table 4.2: The ‘sentential’ algorithm based on the process of constructing an individual description by conjoining terms for each predicate or its negation (“no valid conclusion” is abbreviated NVC).
translate premisses  $A \rightarrow \neg B, B \& C$

find source premiss  $B \& C$

$B, C$  conjunction elimination

$B, A \rightarrow \neg B$

$\neg A$  modus tollens

$\neg A, B, C$

make conclusion  Some Cs are not As

Figure 4.3: ND method used to solve the syllogism "No As are Bs, Some Bs are Cs."

- find the individual: describe the individual in terms of the properties by applying either the rule conjunction elimination or modus ponens to the source premiss;

- find out about the individual: use modus ponens or modus tollens on the second premiss in order to describe the individual in terms of the third property, if possible;

- make a conclusion: refer to the forms of the premisses in order to judge whether the description of the individual is general or particular, i.e., deduce whether the conclusion is universal or existential.

The syllogism "No As are Bs, Some Bs are Cs" is solved using the ND method in Figure 4.3, compare this to Figure 4.1 indicating the EC method used to solve this syllogism.

4.3.3 Similarities and differences between the EC and ND methods

The similarity of the two methods is based on their implementing the same underlying algorithm. Both methods rely on the case-identifiability of the syllogistic system, and both methods require the identification of critical individuals and then a quantification of the individual to form a conclusion.

Similarities between the representations of the premisses are apparent. In the EC method particular premisses are represented by overlapping circles, in the ND method particular premisses are represented by the "&" sign. Universal premisses are represented by complete circles or the "→" sign, respectively. Negation is represented differently in the two methods, this will be discussed below.
Both methods require operations of either *modus ponens* or *modus tollens* when integrating the premisses. Figures 4.4 and 4.5 illustrate these operations in the two methods. *Modus ponens* in

![Figure 4.4: Modus ponens in the EC and ND methods for syllogism “All Bs are As, Some Bs are Cs.”](image)

The EC method is when the third circle in the integrated diagram encompasses the critical region. *Modus tollens* is when the third circle is positioned so that it does not overlap the critical region.

There is a constellation of differences between the two methods, hinted at, but not explored by Stenning and Yule (1997). They note that the ND method focuses on the construction of the individual description whereas the EC method fully determines which sub-regions appear in the final diagram before reading off the conclusion. The ND method is focused on a single x-marked region, whereas the graphical method tests all the critical individuals simultaneously. Stenning and Yule suggest that the cognitive process of implementing the graphical method may assess the several critical regions in the diagram serially, and therefore the distinction may be more apparent than real. Indeed, both methods can be adapted so that individuals are either resolved simultaneously or focused on singly.

The ND method requires that the critical individual that is the focus of reasoning is selected at the outset. When the critical individual is selected, operations are then made on that individual to ex-
press its properties in terms of both premisses. The EC method makes no such requirement on an initial selection of the critical individual before operating to integrate the premiss information: the critical individual is selected after the integrated diagram has been produced. Comparing *modus ponens* and *modus tollens* in the EC and ND methods indicates that for the ND method which rule is implemented is pre-meditated. For the EC method, whether *modus ponens* or *modus tollens* is used is *post hoc*, depending on which individual turns out to be critical. This difference in focus is the principle difference in the way the ICIA is implemented in the two systems. As this difference in focus means the order and type of operation differs, the representational differences are underwritten by algorithmic differences. This account does not conflict with the aims of Stenning and Yule (1997). The difference in process determining the two algorithms is not unrelated to the properties of the systems of which the EC and the ND systems are fragments. However, the ICIA is seen to be a class of algorithms, each of which is paired with a representation. The contrast will be referred to as a focused or a parallel resolution of critical individuals. Some resonance with the serialist/holist learning style is suggested by this distinction.

The two methods also differ in terms of their representation of negation. Consider the translation of “No As are Bs” in the two methods. In the EC method two individuals are marked by crosses, in the ND method only one individual is available—the individual with property A. To discover the other individual directly available in the EC method—the individual with property B—the ND representation must be operated upon: contraposition\(^2\) gives \(B \rightarrow \neg A\). In the ND method negation is propositional, so that the negated properties \(\neg A, \neg B, \neg C\) are represented explicitly. In the EC method, negation is sentential so the negated properties are left implicit in the background of the diagram (everything outside the A circle is \(\neg A\)). The EC method can be adapted so that circles represent \(\neg A\) properties explicitly, and this will then be a focused resolution implementation of the ICIA. Thus the differences between algorithms in the ICIA class is modality-independent.

Consider Figure 4.6 which shows the EC method with propositional negation used to solve the syllogism “No As are Bs, Some Bs are Cs”.

In line with the ND method solution for this syllogism (Figure 4.3) the representation of the premisses must be changed before the critical individual to be operated upon can be isolated—in this case, the individual in the B circle. This adapted EC version removes the represented symmetry of the “no” premiss that is found in the EC method. Recognising the symmetry of the

\(^2\)along with double negation elimination.
No As are Bs

Some Bs are Cs

**conclusion:**
Some Cs are not As

Figure 4.6: Adapted EC method used to solve "No As are Bs, Some Bs are Cs."

"no" relation is a rich source of individual differences in syllogistic reasoning, and is discussed below.

Bird (1964) explored extending systems of syllogistic logic by introducing propositional negation, so including two additional logical forms: "all not A are B", and "some not A are not B". This is precisely the system that is expressible if negation is explicit in the representations. Without any extension to the language, the ND method can be extended to include these forms: \( \neg A \rightarrow B \), \( \neg A \land \neg B \). In the EC method, whereas "Some not As are not Bs" can be expressed with a simple extension to the system, the only ways to extend the system to include a representation of "All not As are Bs" are either to extend the system to represent the universe, or, as above, to extend the system to use explicit, or propositional, negation. Hence, the ND method draws tokens from a more expressive language fragment. The EC method draws on a symbolic system that more closely respects the constraints of the syllogistic system. This constraint on the representation of negation is respected in the conventional first order logic expression of the syllogistic premisses (see Table 4.3), and perhaps this is the reason for the manner of representation that has been preferred.

The sequential/parallel and explicit/implicit negation properties of the algorithms underlying

---

3 There is a long tradition of considering negative terms in syllogistic premisses, the Medieval logicians considered them at length: Pope Leo XXI, in his *Logica Grammaticus*, provides an infallible guide. However, their use in psychology of reasoning research is, as far as I am aware, non-existent.

4 See Stenning & Yule, 1997, for discussion of syllogisms with this "U-conclusion."
All A are B \quad \forall x (A x \rightarrow B x)  
No A are B \quad \neg \exists x (A x \land B x)  
Some A are B \quad \exists x (A x \land B x)  
Some A are not B \quad \neg \forall x (A x \rightarrow B x)  

Table 4.3: The conventional representation of syllogistic premisses in first order logic.

the EC and ND methods are related. It is because the focus is on one individual that information about the individual is fully specified in the ND method (i.e., negation is propositional). Information about individuals is not always explicit in the EC method – only after resolution of the critical individuals might negated properties be realised – and this means that more individuals can be resolved simultaneously. Schroenens et al. (submitted) considered a phenomenon closely related to the propositional/explicit and sentential/implicit negation distinction when he attempted to distinguish between “implicit affirmation” (e.g., “A” affirms “not-B”) and “implicit denial” (e.g., “A” denies “B”) in a propositional reasoning experiment.

The distinction in properties between the algorithms highlights a dual confusion inherent in the mental models literature. The first is over how negation is to be expressed in mental models. The second is how and when multiple models are constructed.

Johnson-Laird & Byrne (1991) describe the initial models that subjects build for “No” and “Some...not” premisses as representing negation implicitly:

No As are Bs \quad Some As are not Bs  
\begin{array}{c|c}
[A] & A \\
[A] & A \\
[B] & A & [B] \\
[B] & [B] \\
\end{array}

where squared brackets indicate that the individuals having that property are exhausted. However, models may have to be “fleshed out” by making more individuals explicit, and indicating negation explicitly with a “propositional-like tag” (Johnson-Laird & Byrne, 1991). Fully explicit models of the two negative premisses look like this:
If a fully explicit model is first constructed, then problems will not require more than one model for solution. The unitary model will, however, have many individuals in it. This relates to a criticism of mental models theory by Ford (1995), where she indicates that representing "Some As are not Bs" like this:

\[
\begin{array}{cc}
A & -B \\
A & -B \\
A & B \\
A & B \\
\end{array}
\]

makes syllogisms with valid conclusion containing this premiss one-model problems instead of two-model which are required if the Johnson-Laird and Byrne (1991) representation is used. Hence, if the "right" model is selected initially and, concomitantly, negation is made explicit, the number of models is reduced to one.

This means either that mental models is a sufficiently powerful framework to incorporate the difference between the sequential/parallel algorithms in subjects' reasoning, or that it is weak due to underspecification of the implementation used. Determining the initial model of the premisses that subjects construct may be a useful means of identifying which algorithm the subject is operating. A recent study by Roberts (submitted) on relational reasoning problems lends some credence to this processing distinction. Subjects can be distinguished into those that construct only one model for relational reasoning problems (e.g., The fork is to the left of the knife. The spoon is above the knife. Where is the spoon with respect to the fork?) and those that construct more than one model. This seems to be a robust characteristic of subjects regardless of the problem's characteristics, i.e., whether it is a single-model, a multiple-model, or an indeterminate problem.

A further source of evidence for the different treatment of negation as a symptom of the alternative processes underlying representation selection is whether the symmetry of the no premiss
is observed by the subject. The way subjects interpret quantified statements may provide a potential tool for diagnosing representation use. Of particular interest are the negative quantifiers "No", and "Some...not", as the previous discussion suggests interpretations of these will vary according to whether negation is treated propositionally or sententially. Subjects that like to focus on one individual, and explicate the description of that individual are more likely to use propositional negation, as this enables the properties of the individual to be expressed explicitly. The alternative interpretation – sentential negation – will be the hallmark of subjects more likely to use the parallel resolution algorithm for solving syllogisms.

Newstead (1989; 1995) devised an "immediate inference" (II) task in order to analyse the extent to which errors in syllogistic reasoning could be predicted from errors in interpretation of the premises. He examined two issues of interpretation: illicit conversion errors (for example, erroneously converting "All As are Bs" to "All Bs are As") and Gricean implicature, where errors are made according to the Maxim of Quantity (interpreting "All As are Bs" as not implying that "Some As are Bs" is true, for example). The II task is a questionnaire consisting of four pages each with one of the four standard quantified statements (All As are Bs; Some As are Bs; No As are Bs; Some As are not Bs) at the top of the page. Beneath this statement, the four quantified statements (e.g., All As are Bs) and their converses (e.g., All Bs are As) were displayed. In one version of the 1995 test, there are three options for subjects to mark next to each statement: "definitely true", "definitely false", or "possibly true, possibly false". Subjects were instructed that "Some" does not preclude "All" and that "Some...not" does not preclude "No".

Newstead found that both types of error (conversion and Gricean) were common, but they failed to predict errors in syllogistic reasoning. His explanation for this lack of prediction is due to the greater "depth of processing" in syllogistic reasoning, and as a result interpretation errors do not commute to this task. His conclusions are unsatisfactory for a number of reasons. First, his model of syllogistic reasoning is tied to a standard version of mental models theory (Johnson-Laird & Byrne, 1991) and so issues of strategic variation and style are not considered. This chapter has argued for the importance of such considerations when analysing reasoning. Second, the depth of processing account seems unsatisfactory as errors of conversion, for example, can occur in syllogistic reasoning and not in the II task. A student in a small-scale study made no conversion errors for "All As are Bs" in the II task, but then reasoned as follows for the syllogism "All As are Bs, All Bs are Cs":

89
All As are Bs ... and All Bs are Cs, all As are Bs then all Bs are also As, so all As are also Cs.

The student was then asked if it is true that all As are Bs means all Bs are As, and she responds “not necessarily”. Then she was shown her response to the II task where she had indicated “can’t tell” and she accepts this as true. However, when she is then presented with the syllogism “All Bs are As, All Bs are Cs” she again makes the conversion error (problems were presented as people taking courses in architecture, botany and chemistry at a party):

If all botanist are architects and all botanists are chemists ... then all architects have to be chemists ... if all botanists are architects then all architects are botanists ... so ... they all have botany in common so if you spoke to a chemist and they said they were a botanist all botanists are architects so all chemists must be architects.

This dialogue suggests that depth of processing is not greater in the syllogistic task, in fact it is perhaps the opposite that is true for this student, as she concedes that “it depends how simply you look at each category” as to whether conversion of “All As are Bs” is valid. Later, despairingly, she acknowledges the effort of the task, uttering: “oh my God I have to use my brain!” This does not preclude the use of the II task in exposing processes of reasoning, but it does argue against depth of processing accounts that omit mention of stylistic preference. It seems to be the case that this student varies her interpretation of the premisses, and attempting to understand the conditions and strategies influencing this interpretation shift are a topic for individual difference research.

A third problem with Newstead’s account is that Stenning and Cox (1995) have shown that quantifier interpretation is a matter of stylistic variation and these styles relate to responses in syllogistic reasoning. Stenning and Cox used the same test as Newstead (1995), where available responses to each statement were “true”, “false”, or “can’t tell”. Subjects were told to treat “some” as meaning “at least one and possibly all”. They noted general patterns of response to the statements, distinguished in terms of whether the subject responded “rashly” by stating “true” or “false” when “can’t tell” is correct, or “hesitantly” by responding “can’t tell” when “true” or “false” was correct. Further, these responses were also determined by whether subject-predicate order was the same in the given statement as in the target statement (in-place) or whether it was reversed (out-of-place). Responses by subjects were seen to fall neatly into this two-way
grouping (rash/hesitant by in-place/out-of-place), enabling a classification of response that is independent of particular quantifiers. Several subjects seemed to react consistently according to the grammar of the statements more than they did to the individual quantifiers, so whether the statement was in-place or out-of-place took precedence over the individual quantifiers.

Stenning, Yule and Cox (1996) related responses in the II task to syllogistic reasoning, where all 64 syllogisms were given to subjects. They found that subjects who responded rashly on the II task made more “no valid conclusion” responses to syllogisms. Also, subjects who made rash in-place errors but did not make hesitant out-of-place errors demonstrated the least figural effect: the order of terms (As and Cs) in their conclusion was not influenced by the order of terms in the premisses. Subjects who did not make rash in-place errors but made hesitant out-of-place errors exhibited the strongest figural effect: the effect of grammar on these subjects was greatest and was reflected in their II response profile. Subjects who tend not to switch subject and predicate are those that are highly influenced by the order of premisses in syllogisms when making a response.

Of these hesitant/rash in-place/out-of-place responses, those involving the negative premisses are especially interesting for reasons discussed above. So, do subjects accept the conversion of the “No” premiss, i.e., given “No As are Bs” is true, is it true, false, or not possible to tell whether “No Bs are As”? Not accepting this “no-conversion” is an instance of a hesitant out-of-place response, and seems to be a good predictor of this tendency. Subjects that respond “yes” to “no-conversion” tend to make fewer hesitant out-of-place responses than those that respond “can’t tell” (from a population of their 20 University of Edinburgh students, mean hesitant out-of-place scores were 0.45 for “no-conversion” students compared to 6.78 for those that did not accept the conversion, t(18) = 7.69, p<0.001). This question also reflects the extent to which critical individuals are left implicit or are specified in the algorithm used by the subject. Rejecting no-conversion means the individual A¬B is accepted, but the individual B¬A is not specified. Accepting no-conversion means that both these individuals are specified. The latter is a hallmark of the focused approach where individuals are specified, the former suggests the parallel resolution algorithm.

No-conversion suggests itself as a useful reflection of general tendencies to response in syllogisms and also as a predictor of use of the different algorithms in reasoning. The connection between the theoretical methods for solving syllogisms and the empirical strategies of Ford (1995) are next discussed to provide some justification for the existence of the different algorithms in spontaneously developed approaches to problem solving. After this brief foray, the generality of
the differences between the sequential and parallel resolution algorithms is explored, linking the properties of the algorithms to the strategies used in HP.

4.4 Linking Ford's strategies to the EC/ND methods

Rewriting Ford's verbal strategy principle in the representations used in the ND method enables a closer comparison between the methods. This translation is shown in Table 4.4.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Ford's Verbal Strategy</th>
<th>ND Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>Given B → A and C &amp; B, assume C and B and infer A.</td>
<td>A.(i)</td>
</tr>
<tr>
<td></td>
<td>Given B → A and C → B, assume C, and infer B and A.</td>
<td>A.(ii)</td>
</tr>
<tr>
<td></td>
<td>Given A → B and C → ¬B, assume C and infer ¬B and ¬A.</td>
<td>B.(i)</td>
</tr>
</tbody>
</table>

Table 4.4: Ford's verbal strategy rewritten in the ND terminology.

The substitution feature of Ford's verbal strategy relates to finding the source premiss. In Ford's terms, the source premiss is the premiss providing the term to be substituted. Then, the ND and Ford's verbal strategy are similar in applying either *modus ponens* or *modus tollens* to the other premiss – the substitution-in process is the same as the procedure of finding out about the individual. This means that direct mappings can be made between using simple or sophisticated substitution and the rules in the ND method for finding the source premiss, and between use of rules A.(i), A.(ii), B.(i), and B.(ii) in Ford's strategy and use of *modus ponens* or *modus tollens* in the ND method. In short, Ford's verbal strategy is a focused critical individual implementation of the ICIA.

There are two differences between Ford's verbal strategy and the ND method. The first is that the verbal strategy does not provide a procedure for reaching the conclusion, rather it provides the principles that Ford perceived to be those subjects were striving to adhere to. The ND method, in contrast, does provide a decision procedure for gaining the solution. The second difference is that the verbal strategy is affected to a large extent by premiss order during the substitution choosing
phase (selecting the source premiss), whereas the ND method abstracts from term order in the premisses. So, for Ford’s verbal method, principle A(i) will be more difficult to apply when the syllogism is “All Bs are As, Some Bs are Cs” than if it is “All Bs are As, Some Cs are Bs” as then the second premiss must be re-ordered before the inference or substitution can be made, and this is supported by the data from Ford’s experiment. This does not mean that such ordering effects do not emerge in applying the ND method, indeed it is anticipated that there will be such effects during teaching: the influence of ordering seems to be a hallmark of verbal strategies in general.

The algorithm underlying the ND method together with requiring an explicit description of each individual at a time has this sequential property. In Stenning and Yule’s (1997) framework, these would be precisely the properties of the larger system of which the ND method is a fragment that the method would inherit.

The spatial strategy and the EC method bear close resemblance, particularly as some of Ford’s subjects attempt to devise the cross-notation for their set diagrams. Ford’s spatial strategy requires operations on several individuals at once, represented in the relationships between circles by the subjects. Hence, like the EC method, it is an implementation of the ICIA where all individuals are resolved simultaneously. Ford’s spatial strategy and the EC method are also alike in that neither predict large effects of term order in the premisses, or of surface features of information presentation.

### 4.5 Strategies for making reasoning tractable

Ford’s strategies indicate that the differences between the algorithms implemented by the EC and ND methods are realised spontaneously without training in any method. The differences between the algorithms can be described in terms of whether individuals are resolved sequentially or in parallel. Parallel resolution obviates the need for full specification of individuals, sequential resolution requires fuller specifications of each individual which in turn requires propositional negation. The holist/serialist distinction in tandem with this characterisation in terms of specifying properties of individuals suggests a connection with the strategies observed in HP in the previous chapter. There, exhaustive cases are constructed either by specifying each case (serialist proof) or by constructing nested cases that contain more abstraction (holist proof). The extent to which information is specified in the representation discriminates HP strategies as well as syllo-
gism algorithms. This distinction was developed to describe the alternative methods of solving problems in the reasoning tasks discussed in Chapter 2, but HP and syllogisms are domains where each approach proves to be equally effective. Therefore, styles rather than abilities for representing abstract information are being exhibited.

Where do these preferences come from? One potential exploration of origins comes from similarities between the strategy - representation - algorithm account above and Levesque’s analysis of different strategies for making reasoning tractable. Humans are faced with a rather difficult computational task – to reason in real time with very large databases of information. Logic, according to Levesque, even if it was good psychology, is too demanding computationally. So, what is needed is some description of how humans solve tasks where the space of possible inferences and cases to be considered is immense. One way to approach an answer to this question is to look at the conditions on knowledge-based systems such that they can compute in real-time.

Of the five alternatives for making reasoning tractable discussed by Levesque (1988), two stand out as reflecting the approaches exemplified in the EC and ND methods. The first strategy for making reasoning tractable is to make information vivid. The other method is to limit the number of cases that are considered. These correspond to making each case fully specified, or by considering fewer cases that are not fully specified but that cover the space of possibilities. Levesque phrases the distinction thus:

What does this have to do with tractability? The point is this: The more that is left unsaid, the more possibilities are allowed by what is said. To determine what is entailed by what is said, all of these possibilities have to be covered one way or another. Proof methods may differ on how they do this, but none of them (have been shown to) do substantially better in the limit than a case analysis of all these possibilities. The problem is that cases do not simply add up, they multiply: with n independent binary choices, there are 2^n cases to consider, too large for all but very small n.

This suggests one way to keep reasoning tractable: arrange the task so that the number of cases that have to be (conceptually) considered is kept manageable (Levesque, 1988, pp.370-1).

Representing information vividly means that each piece of information is expressed as a “ground, function-free atomic sentence” (Levesque, 1988, p.371). To illustrate the approach, Levesque discusses the representation of negation. A vivid representation of negation “p is not A” means that p has the property of being non-A. If “Jane is not married to Jim” is represented vividly, then Jane and Jim have the property of being not-married. Coupling this with the information “Jane is mar-
ried to John or Jim" will not render the conclusion "Jane is married to John" in a vivid database, because the contradiction between being married and being not-married is lost. Levesque found that reasoning with vivid representations of negation corresponds precisely to the inferences that are licensed by relevance logic: a logic of implicit belief (Levesque, 1984). A database represented this way has the property that entailment becomes a table lookup process, making very large knowledge bases tractable. The algorithm implemented by the ND representations uses vivid information, whereas that implemented by the EC method does not.

An alternative method of making reasoning tractable, discussed by Levesque, is to limit the number of cases from the database for consideration. The algorithm implemented by the EC method only makes available three cases – A, B and C – whereas the ND method has up to six cases permissible, having in addition the three negated properties ∼A, ∼B, and ∼C. Keeping the number of cases to a manageable number requires that the cases are not so specific so that they cover the space of possibilities. This is especially the case in the HP strategies: a higher level of abstraction means fewer cases have to be constructed, but the cost of this is that "more is left unsaid" and so consequently "more possibilities are allowed by what is said." For the holist proof in HP, use of a "Merge" step after an exhaustive cases structure has to select between several situations consistent with the set of abstract situations, whereas for the serialist proof the "Merge" step is far more constrained.

The claim is that humans use one or other of these strategies to make their reasoning tractable, and these are implemented as algorithms that either specify several cases but do so sequentially, or express cases with more abstraction in parallel. The general approach to making reasoning tractable is manifested in particular tasks, even if the database is not unmanageably large in these constrained domains.

The preferences for the sequential/specific or the parallel/abstract approach to making reasoning tractable are hypothesised to be reflected by the psychometric measures discussed in the previous chapter. The precise hypotheses connecting the serialist/holist dimension, the FID dimension, the GREA test, and response to "no-conversion" with respect to experiments testing learning and use of the different algorithms, representations or strategies are considered in the next chapter.

The next section reviews previous studies of teaching methods for solving syllogisms. This is included as background for the teaching syllogisms study to be presented in Chapter 6.
4.6 Previous studies of teaching syllogisms

There have been several studies that have taught subjects to solve syllogisms. Some have offered graphical representations to support reasoning, others have used some system of verbal rules. Not one of these studies has probed in detail the responses of subjects to the problems, and have instead looked at overall acquisition of the method. Also, no studies have directly compared teaching with graphics and with sentential representations in an aptitude-treatment interaction study. The relationship between subject and strategy remains underspecified from these studies. The teaching study detailed in the next chapter aims to expand on these studies. The remainder of this chapter serves as a review of the extant teaching syllogism literature.

Frandsen and Holder (1969) pre-tested subjects on a Verbal Reasoning and a Space Relations test taken from the Differential Aptitude Test. 18 pairs of subjects were then matched for being identical in Verbal Reasoning score, but as disparate as possible in Space Relations score. Subjects were then placed into either an instructed group or a control group. All subjects were given a test made up of verbal problems, including syllogisms, then one group of subjects were given instruction for approximately 50 minutes in graphical methods for solving these problems. For syllogisms, subjects were instructed in Venn diagrams. Then in a post-test another set of verbal problems were presented of the same type as in the pre-test. They found that for subjects who scored low on the Space Relations test, those that were instructed improved significantly more than those who did not receive instruction (t = 2.04, p<0.05). Also, these instructed low Space Relations subjects improved their score significantly more than did instructed subjects who scored high on the Space Relations test (t = 3.06, p<0.01), however the high scoring subjects tended to gain close to maximum score on the pre-test, so this group is subject to ceiling effects in the test.

Frandsen and Holder's study shows that training in graphical methods helps some subjects to solve verbal problems, including syllogisms. However, the type of syllogism used is unspecified, and whether syllogism solution in particular benefitted from instruction is unclear, as each type of verbal problem is not distinguished in the results. Syllogisms made up only four of the fifteen verbal problems in each of the pre- and post-tests. This study certainly leaves room for more specific studies of teaching and aptitude in syllogistic problem solving.

A further study that connected syllogistic teaching to aptitudes of subjects was conducted by de Leeuw (1980). Two graphical methods for solving syllogisms were taught, one based on Venn
diagrams provided a decision procedure for solving the problems, the other, based on Euler's circles, encouraged generation of counterexample diagrams, but did not provide a step-by-step method, but instead “a process of divergent thinking is required” (p.29). 30 tenth grade school subjects participated. Problems were presented with conclusions, and the task was to assess whether the conclusion was valid or invalid given the premisses.

The Venn diagram method (described as “algorithmic”) represented the syllogism with three overlapping circles, with areas shaded to indicate parts of the diagram that are empty regions, and a cross to indicate non-empty areas. In the instructional program, the subject was taught to operate five subalgorithms:

- Draw the base diagram and label the circles;
- represent the first premiss in the diagram (this has to be a universal premiss, if there is one);
- represent the second premiss;
- decide whether a + could be placed in the diagram (this has to be placed at the border of the two regions unless one of the regions is crossed out)⁵;
- decide whether the conclusion is valid.

The Euler's circles method (the “heuristic” method) instructed subjects in drawing all the possible relationships for the two premisses (for the selected syllogisms there was a maximum of five diagrams and a minimum of one diagram possible). Then subjects were taught to look for a diagram that contradicted the given conclusion. If this was possible, then the conclusion is invalid, if search failed, then the conclusion is valid.

Subjects were pre-tested on a syllogism task composed of problems to be used in the instructional programs; the embedded figures test (Witkin et al., 1971); an embedded syllogism test, where syllogistic problems were phrased in paragraph form; and syllogisms with three premisses (sorites). Delayed post tests (2 to 4 days after the end of the instruction) were also presented. These involved the instructional syllogisms, the embedded syllogisms, and the sorites. These post-tests were also given to subjects two months after instruction as tests of retention.

⁵Note that this placing of crosses at the border between regions was attempted by several subjects in Ford's (1995) study.
De Leeuw found that training was fairly successful, with Venn-trained subjects getting 69.3% of the instructional problems correct in the post-test, and Euler-trained subjects getting 66.7% correct. Two thirds of errors for the algorithmic method were caused by incorrect assessment of correctly drawn Venn diagrams. Using the algorithmic method increased performance from pre-test to post-test for universal conclusion problems (+0.5), and slightly decreased performance on particular conclusion problems (-0.1). Venn-trained subjects performed better on universal conclusion problems than Euler-trained subjects (0.9 compared to 0.5), but did worse on the particular conclusion problems (0.6 compared to 0.8): F(1, 28) = 12.83, p = 0.001. For the delayed instructional syllogism post-test, the Euler method produced better performance (F(1, 28) = 8.01, p=0.009). For the sorites in the post-test, the Euler group performed better than the Venn group (F(1, 28) = 4.48, p=0.043). The EFT related to performance on the post-test embedded syllogism task (F(2, 27) = 8.51, p=0.007). Transfer was greater for those subjects that scored high on the EFT. Euler-trained subjects tended to use diagrams at a consistent level on the post-test and delayed post test problems (31% and 29%, compared to Venn subjects' 35% and 2%, respectively).

De Leeuw's (1980) study shows that two graphical methods are distinct in terms of the strategies that they promote, and that these strategic differences provoke differences in subjects' learning. Problem type is sensitive to these strategic differences, as shown by the different performance of subjects on particular conclusion problems and universal conclusion problems. The relationship between the EFT and performance indicates that psychometric tests have a predictive role in complex task performance. It is perhaps not surprising that the high-scoring EFT subjects' ability is seen only on embedded syllogisms, but it does suggest that the EFT relates to an ability to abstract and apply the taught method as presented in either instructional program. Witkin et al. (1977) found that the field-dependence/independence distinction predicted differences in learning only if the subject matter is unstructured, so it is perhaps surprising that from two well-structured teaching methods a difference was found at all.

Grossen and Carnine (1990) taught learning disabled schoolchildren a method for solving syllogisms that was based on Euler's Circles. Their method required the subject to reason with Euler's Circle diagrams in their standard form. Thus, the cross-marking facility of Stenning and Oberlander's (1995) method was not used and so several arrangements of the circles had to be used by the subject. These diagrams were used in conjunction with rules for which diagrams to select in order to reach a conclusion, and in what order. The premiss form "Some As are not Bs"
was not used, and syllogisms with conclusions of this form also did not appear. There were two treatments: one where subjects had to construct diagrams in collaboration with the computer-based instruction, and one where subjects were not required to construct diagrams themselves. They found that the diagram-construction subjects took the same time to get through the material, but required fewer questions by the instructor (t(24) = 2.17, p=0.035). Subjects were divided into those that were proficient on using the Euler's Circles and those that weren't. The proficient group scored significantly higher on a transfer task that required the subject to solve sorites (Wilk’s lambda = 0.38, p=0.008).

Grossen and Carnine's (1990) study shows that subjects learn a graphical method for solving syllogisms to differing degrees of success. Their analysis also supports use of measures of instructor intervention in order to assess acquisition of the method. Their study does not, however, address issues of learning style, and nor is there any comparison with how subjects may learn from sententially-based materials (indeed, their Euler’s Circles method seems to compound graphical and sentential operations).

Gilhooly, Logie & Wynn (in press) in a study of the role of working memory in reasoning, attempted to teach their subjects a method for assisting with solving syllogisms. This was due to the predominance of strategies that operated on the surface features of the syllogistic problems (atmosphere, matching, or guessing). Such strategies are not affected to any great degree by dual-tasks, whereas subjects who used a “logic” strategy (i.e., those that solved more syllogisms correctly than the other methods predicted) deteriorated in performance or changed their strategy as a result of loading on working memory. As the teaching method was not the principle reason for the study, it is only discussed briefly by the authors:

The general method of solution ... was to try to convert problems into the form S-M, M-P (i.e., Figure 1) and then to reason in terms of set membership. To achieve the desired Figure, premises may need to be re-ordered by conversion, e.g., from “All P are M” to “Some M are P” (Gilhooly et al., in press, p.21).

The method invokes aspects of the identify critical individuals algorithm – converting the premises to Figure 1 is akin to the process of selecting the source premiss in the ND method for many problems. Reasoning about set membership leaves open the possibility of spatial or verbal operations, so it is possible that subjects in their study either used a sequential/specific or a parallel/abstract approach to the problems. Subjects were trained by solving 20 syllogisms which,
if they got one wrong, had the solution for that problem explained to them verbally. The trained subjects scored higher on another set of 20 syllogisms than a control group: 11.39 compared to 3.83 for the untrained group (t(32) = 54.80, p<0.00001).

Bauer and Johnson-Laird (1993) instructed subjects in using diagrams in order to solve disjunction problems of the form:

Raphael is in Tacoma or Julia is in Atlanta or both.
Julia is in Seattle or Paul is in Philadelphia or both.

What follows?

This study is included here as it raises issues concerning the use of graphics to assist with reasoning, and the importance of finding a “good” representation. Also, disjunctive reasoning has been closely compared to syllogistic reasoning by mental models theorists (Johnson-Laird & Byrne, 1991) who argue that this form of reasoning relies on the same processes as syllogistic problem solving.

Bauer and Johnson-Laird’s study compared subjects’ conclusions when premisses were represented graphically to when they were presented just verbally. The diagrams used boxes linked by lines either with a circle to indicate inclusive disjunction or a square to indicate exclusive disjunction. No difference was found between graphical and verbal presentations. They concluded that this was because “the diagrams failed to make sufficiently explicit the alternative states of affairs” (p.373). In addition, the inclusive/exclusive disjunction distinction was depicted in an arbitrary fashion. Their second experiment remedied this by representing disjunctions using diagrams reminiscent of electronic circuit diagrams. For diagram representations, subjects drew 74% correct conclusions, compared to 46% correct for the verbal problems. Subjects also responded faster to the diagrams (F(1, 44) = 8.59, p<0.01). The authors conclude that the diagrams help subjects by making all the possibilities explicit. But this does not determine whether the subjects’ initial difficulties were because all the individuals to be specified were not exhausted with a sequential/specific approach, or because the individuals were not explicit enough (and so a problem with the parallel/abstract approach). A graphical representation may assist in either approach, and so the data from task performance has to be more detailed in order to settle this issue.
4.7 Summary of chapter

The literature reviews in this chapter indicate the fruitfulness of syllogisms for examining issues of strategy, representation, algorithms and connections between accounts using these terms. Syllogisms are a domain where psychometric and computational accounts are complementary. The reviews also consolidate an approach that deals with stages in representation use. Lastly, a review of teaching syllogism studies indicates the paucity of such studies, particularly absent are controlled comparisons of teaching using different representations and assessment of individual difference variations.

The theoretical work shows that the EC and ND methods developed by Stenning and Oberlander (1995) and Stenning and Yule (1997) can be distinguished in terms of the implementation of the ICIA: whether they are a sequential/specific or parallel/abstract approach. These differences are symptomatic of the larger representational systems that the fragments used in the methods are drawn from. This is a theme further explored in Chapter 8 when theories of thinking and theories of reasoning are discussed.

Furthermore, distinguishing methods in terms of strategy, representation and algorithms highlights and resolves confusions over formulations of the mental models theory and captures the differences between several theories of internal representation in reasoning. Such a framework allows the generation of testable hypotheses that enable an empirical distinction of syllogistic reasoning in terms of preferences for certain representations. These hypotheses for the teaching studies are presented in detail in the next chapter.
Chapter 5

Aims of the empirical studies

5.1 General aims of the empirical studies

The studies on strategy use in reasoning suggest that there are differences in the use of abstraction in representation, and that this is one way of classifying the different strategies used. Ability to develop a more appropriate level of abstraction is related to scoring high on spatial ability measures. The HP studies show that use of abstract or concrete representations is a matter of style as well as ability: some subjects stick to using abstract representations, even when this is not optimal for the task. One main purpose of the studies to be conducted is to distinguish styles from abilities for subjects' selection and use of different representations. This will be done by relating psychometrics to performance on syllogisms and on HP. The syllogism study reported in Chapter 6 compares students' learning to solve syllogisms using either the EC or the ND method. Students are grouped according to their scores on the psychometrics discussed in Chapter 3. The HP experiment, reported in Chapter 7, relates students' learning from and learning within HP to their performance on psychometric measures.

The psychometrics will be interpreted from the perspective of propensities and preferences for using different levels of abstraction, and their relation to well-defined tasks will have the subsidiary aim of illuminating what it is they are testing. Psychometrics are used to discriminate performance, but are also a topic of investigation themselves.

The spontaneous use of different strategies in categorial syllogisms suggests that this is a good
domain for studying issues of style. The two methods for solving syllogisms indicate that both are equally appropriate for solving the problems, but vary over the treatment of individuals in constructing descriptions of them. HP reveals this property of different levels of abstraction in representation in common with the methods for solving syllogisms: abstract representation strategies limit the number of cases considered, concrete representation strategies limit the operations that can be performed on the set of cases once constructed.

These links require some support from empirical studies. Relating performance on different methods for solving syllogisms with performance on HP via psychometrics will enable these connections to be tested. Hence, generalising the aptitude-treatment interaction observed in HP to other reasoning domains is an important aim of the studies.

Finally, unpacking the nature of the transfer of skills from the HP course will be undertaken. This requires a reappraisal of the GRE test, and this necessitates the development of a generalised taxonomy of problem types. Relating psychometrics to the transfer of skills particularly associated with different problem types will contribute to the debate on what is learned through following logic courses. In addition such a connection will provide finer-grained analyses of the effects of teaching with different types of representation, representations that vary in terms of the level of abstraction they employ and the different styles of computational process induced by the representations.

This is a summary of the four most general aims of the empirical studies. The next sections of this chapter present hypotheses particular to the syllogism study and the HP study, cashing out specific predictions relating the reasoning domains to the psychometric tests.

5.2 Styles and strategies in syllogisms

The discussion of different implementations of the same task, with the same underlying algorithm enables a more precise comparison of teaching methods than has been available to other investigations of learning problem solving skills. The close analysis of differences in the two methods suggests ways in which the graphical and sentential representations may relate to different strategic performance, and these differences are founded on alternative styles of computational processing.
The similarity of the two methods enables an empirical comparison of the two methods in terms of controlled differences which result from differences in implementation of the logic of the task. In particular:

- The ND method processes each individual sequentially, providing specific descriptions of each individual in terms of properties A, B, and C. The EC method resolves all individuals in parallel and only then provides a full description of the critical individual.

- Relatedly, negation is represented differently in the two methods, either propositionally or sententially and this is a concomitant property of the two styles of processing.

For purposes of analysing responses to the two taught methods, three stages are easily distinguishable and isolable in learning to use the EC and ND methods:

- translating in: translating from the premisses into the represented form;
- manipulating the represented forms in order to isolate the critical individuals;
- translating out of the final representation in order to form a conclusion.

The correspondence of these stages in the two methods mean that different key stages in the use of external representations can be assessed. The differences between the methods discussed above feature differences in manipulating the representations. However, similarities or differences in responses to the other stages in the method illuminate particular features of external representation use and help to specify the nature of strategic differences in using different types of representations in order to support reasoning.

Different responses to the translation in stage will reflect the extent to which the representations promote different styles of processing: the EC translations are conducive to a parallel/abstract implementation of the ICJA, ND representations have more affinity with a sequential/specific implementation.

Differences in response to the manipulating stages hinge on the strategic differences discussed at length above. This stage directly reflects the students' propensity for either sequential or parallel resolution implementations.

Finally, differences in response to the methods in the translation out stage again reflect the students' reaction to representations that reflect sequential or parallel processing. Such differences
will also reflect the extent to which students maintain a trace of the form of the premisses during reasoning: the ND method requires reference to the initial form of the premisses in order to judge whether the conclusion is particular or universal. The EC method requires no such anaphoric reference.

Justification for examining different stages in syllogistic reasoning can be drawn from the literature. Fisher's (1981) approach distinguishes performance on syllogisms in terms of (i) interpreting or encoding the premisses (translating into some sort of mental models representation); (ii) processing models of these interpretations (working memory limitations); and (iii) what Fisher terms "the operation of the deductive strategy" (p.496). The first stage has been the focus of Newstead's (1989; 1995) examination of Gricean errors in interpreting the premisses. Fisher examined the information processing stage by providing students with syllogisms to which they had to mark a series of conclusions as true, false or possibly true (this is the manner of presentation of syllogisms prior to Johnson-Laird and Steedman's (1978) study). Difficult problems are those where many individuals are compatible with the combination of premisses. However, certain types of error were not compatible with processing resource difficulties (they cite as an example subjects' responses to the syllogism "No Bs are As, No Cs are Bs" saying "No As are Cs" is necessarily false). Fisher attributes this sort of error to the final stage: errors in applying the deductive strategy, which is considered to be the process of construction and assessment of mental models (Johnson-Laird & Steedman, 1978). This is because the construction of models to show that "No As are Cs" does not exceed the resources of the student: two models are sufficient, and used by these students for some of the other problems. It is worth mentioning, however, that such responses could equally be explained in terms of matching effects (Wetherick, 1989).

In the framework adopted in the current study, Fisher's interpretation stage matches the translation in stage, and the other two stages map onto the manipulating stage of reasoning, though perhaps errors in using a deductive strategy also incorporates errors in translating out of the representation.

Polk and Newell (1988; 1995) have modelled syllogistic reasoning using SOAR (Newell, 1990). Their model uses repeated encodings of verbal information, and can be interpreted as a staged investigation of problem solving. Their model, termed VR for "verbal reasoning", initially encodes a model of the first premiss, and then adds the information from the second premiss to this model, encoding only the "direct" knowledge from the premisses (i.e., the model respects
the order of terms from the premiss). Following this translation stage, the algorithm generates conclusions based on the ordering of properties in the model. If no conclusion that is legal is generated (i.e., no conclusion that links the end-terms), then re-encoding of the premisses about one property occurs, and a new model is built. The property chosen is the most recently selected property from the premiss encoding. Different versions of VR have been built, which vary in terms of what types of re-encoding they permit. VR1 only encodes direct knowledge, no re-encoding is permitted. VR2 allows the premisses "Some X are Y" to be re-encoded as "Some Y are X", and "No X are Y" as "No Y are X". A third version, allows re-encodings of other premiss forms, including invalid transformations, such as "All X are Y" becoming "All Y are X" and "Some X are not Y" becoming "Some Y are not X". So, in terms of a staged investigation, the VR models all have some distinction between translation in, manipulation and translation out, though the process is not unidirectional.

The different models (VR1 and VR2 – I exempt VR3 due to its invalid re-encodings) fit the framework of theories resulting from the distinction between the properties of the implementation employed in the EC method and that in the ND method, i.e., when and how the critical individuals are selected. VR1 and VR2 both require some identification of the critical individual to begin with, but then VR1 will reach an impasse if the grammar of the problem does not provide a conclusion (this is akin to simple substitution in Ford's verbal method). VR2, on not reaching a conclusion, will re-encode the premisses. If a conclusion is then reached then this is akin to sophisticated substitution. It may be the case that certain re-encodings actually produce models where more than one critical individual is represented, but for a majority of syllogisms, the process is close to the sequential/specific class of implementations (note further that negation is propositional in the VR models).

All the studies previously discussed have been directed towards discovering the stages of internal problem solving with syllogisms. The studies to be discussed will examine stages in the use of external representations, which means that these stages are much more apparent to the experimenter. The relationship between internal and external representation will be addressed later, when the results of the syllogistic and Hyperproof experiments are considered in tandem.

These benefits of the syllogistic domain for studying individual differences in reasoning (controlled implementational differences, stages in representation use, and externalised representations) mean that more detailed features of what the psychometrics are testing can be revealed.
The GREA subscale has been seen to relate to different use of representations in HP. The syllogism replication of the aptitude-treatment interaction hinges on the use of this test in both domains. GREA-Hi students should find the EC method easier, GREA-Lo students should utilise the ND method with greater ease. Using few, abstract cases is consonant with parallel implementations and so GREA-Hi students, as they produce such proofs in HP, ought to prefer a method that implements the parallel/abstract approach to syllogism solution. GREA-Lo students should demonstrate the reverse tendency, as their preference is for generating specific cases. If this relation holds, then it suggests that stylistic influences are at least as important as ability in using abstract representations. There is, however, still the question of whether GREA subscale relates to flexible use of different levels of abstraction. As the syllogism study does not require students to learn various representations, this question will not be addressed, though the HP study will consider this issue directly.

The relationship between scores on the GREV subscale and reasoning domains are little explored. GREV-Hi students benefit from a logic course where the form of representation is uniformly abstract (Stenning, Cox & Oberlander, 1995). However, GREV-Lo students do not seem to respond well to any teaching method, so unlike the GREA subscale it does not seem to be a style measure. The GREV test is included in the syllogism study in order to fully replicate the HP study in another reasoning domain, and also in order to attempt to illuminate what the test is assessing. Again, the HP study, with a generalised taxonomy of problem types, will assist in clarifying the measure.

The serialist/holist learning style has been predicted to be a reflection of sequential or parallel implementation preference: serialist students should find the ND method easier, and holist students should find the EC method easier. Such a finding would support the relationship between representational differences and computational styles. There is a slight tension between accounts in terms of the analytic/holistic and serialist/holist styles with regard to the two "computational" styles that distinguish the different algorithm types for solving syllogisms. Both the analytic and serialist approaches are sequential, but the analytic approach tends to use more abstract representations for solving spatial ability tasks, whereas a serialist strategy will produce concrete situations or hypotheses. So the analytic method presents as a sequential/abstract method, whereas the serialist method is sequential/specific. This is perhaps a difficulty with being able to characterise the nature of the representations used during spatial ability task solution, though. Exactly
how concrete is the holist's representation during a cube-rotation task, for example? Again, this points toward the usefulness of external representation as a window on cognition, it also further indicates the difficulties there are in stipulating performance in psychometric tests. Relating spatial ability test performance to the serialist/holist dimension and performance on the two syllogism teaching methods will enable some resolution of this apparent conflict.

The FID dimension as measured by the HFT will also be measured against learning from the EC and the ND methods. It is anticipated that field-independent students will be better at using more abstract representations to support their reasoning, and field-dependent students will be better at using more concrete representations. This is not at issue in the syllogisms study, as the type of representation is given and fixed in the method. Relatedly, field-independent students were shown not to differ from field-dependent students when the learning procedure was highly structured (de Leeuw, 1980), and so the HFT may not indicate differences between the two methods. However, the extent to which FID maps onto preference for the styles implemented in the EC and ND methods will be indicated by the study.

The relationship between "spatial ability" measures and performance on the EC/ND methods will be tested by using the PFT. Roberts, Gilmore and Wood (1997) found that a "spatial ability" test reflected flexibility in using graphical representations. If spatial ability tests measure flexibility then PFT-Hi students ought to perform better on both the EC and the ND teaching methods. If the PFT is a style measure in the terms defined, then PFT-Hi students will learn one or other of the teaching methods more easily.

The immediate inference (II) task has been proposed as a measure of different treatments of negation that relates to different styles of implementing the ICIA. Subjects who are hesitant out-of-place are those that are least likely to accept the conversion of the "no" premiss. If the relationship between teaching syllogisms and the II task as a gauge of stylistic preference holds, then those that do not accept no-conversion will be more at ease with the EC teaching method that promotes sentential negation as part of the parallel resolution implementation. If students that do accept no-conversion prefer the sequential/specific implementation induced by the ND representations, then this points towards these two strategies as covering the space of possibilities.

Due to similar considerations, students that accept no-conversion are more likely to score high on the GREA test. Such a finding would support the link between parallel/abstract case use as a general strategy in the HP and in the syllogistic domain.
With regard to the other stylistic psychometric measures, it is predicted that no-convertors will have a preference for holist strategies, whereas those that do not convert “no” premisses will be serialists. If the HFT and the PFT relate to the II task then this suggests these measures are stylistic, though there are no specific predictions raised regarding the link between these measures.

5.3 Styles and strategies in HP

The aims for the HP study are to replicate the original aptitude-treatment interactions findings, but also to extend this study in several ways by presenting problems in order to assess the differential effects of varying level of abstraction present in the problems. A further extension and clarification of the original HP study will come from relating performance on HP to the established psychometric measures. The reasoning tasks used to assess transfer of skills from following the HP course can be reclassified in terms of the level of abstraction required by the problems for their effective solution. This level of abstraction can be directly related to that used in the situations constructed by students in HP. Such studies extend the teaching syllogisms study to embrace issues of transfer of skills, the robustness of learned approaches to problems, and the development of representational use.

The HP study will use the same tests as were employed in the original HP study, but performance will also be related to the HFT and the PFT. Unfortunately, the serialist/holist dimension and the II task suggested themselves as useful tools only after the study began, so these will not be used to reflect performance on HP. The range of problems in the exam taken by the students on this course includes a greater range of abstraction than in the original study. This enables a more detailed investigation of the extent of an individual’s variation in representation use. In addition, this broader range of problems means that individual goals can be assessed in the HP exam for the appropriateness of different levels of abstraction in representation, along the lines of the initial investigation into strategy change in HP conducted in Chapter 3. The goal-type investigation is a finer-grained analysis than the problem-level assessment that has been presented in previous studies.

With regard to the psychometrics, it is anticipated that the HFT and the PFT will relate to different use of abstraction in the representations for HP. If the PFT relates to appropriate use of abstraction in representation, then for high-abstract problems a higher degree of abstraction will be used than
for low-abstract problems. If, on the other hand, high PFT scores relate to use of a higher level of abstraction regardless of the problem's requirements, then this will be indicated by response to the HP problems also. The level of abstraction used by students that score high or low on the HFT can also be indicated by situation construction in HP.

Chapter 6 presents the teaching syllogisms experiment. Chapter 7 reports the classification of problem types and the new HP study.
Chapter 6

Teaching methods for solving syllogisms

6.1 Method

6.1.1 Subjects

22 first year undergraduates from the University of Edinburgh participated in the experiment. All subjects were undertaking a first year modular course in Linguistics, which attracts students from a wide variety of academic backgrounds. Subjects were paid £5 for participating in the first session and £8 for participating in each of the last two sessions.

6.1.2 Materials and Procedure

The experiment consisted of three one hour sessions: the first session involved the paper and pencil pre-tests, the GRE test was taken in the second session, and the teaching of syllogisms occupied the final session.

22 students participated in the first session, 21 subjects took the GRE test in the second session, and 17 students participated in the last session.
6.1.3 Pre-tests

As spatial ability has been seen to be an indicator of, or an influence on, learning strategy, a standard measure was included in the pre-tests. The PFT (Ekstrom, French & Harmon, 1976), was used as it has been established that this test demonstrates strategic flexibility in manipulating spatially presented information, as discussed in Chapter 2 (see Figure 2.3). The test is administered in a short time, students having 6 minutes to solve the 20 items.

The Ekstrom et al. HFT was used to measure the FID dimension (see Figure 3.2). The test requires the student to disembed a figure from a background of lines, which is subject to the student’s preference for processing information with or without reference to the context that it appears within.

As Pask’s tasks for measuring the serialist-holist learning style are protracted in their application, a questionnaire-based measure was used. Ford (1984) designed a 16 item study preference questionnaire (SPQ) which required students to state their preference for one of two approaches to a learning situation. An example of a question is shown in Figure 6.1, where subjects mark their preference for one statement over the other by marking a box between the two statements. Ford found that five items were good reflections of strategies in Pask’s Martian animal classification task. Clarke (1993) found that one of the questions in the SPQ was a good indicator of Pask’s learning styles, and this one question is used to discriminate students in the current study. In Figure 6.1, responding with an agreement for the statement on the left indicates holistic tendencies, whereas serialists agree with the right statement. However, if the student did not state a preference for this question then learning style was determined by the five items used by Ford.

The presentation of the SPQ differed from that of Ford and Clarke in two ways:

- Ford and Clarke present their statements in the same order, both within-item and between-item, for each student. In the current study, the order of the items is randomised, as is the order that the statements occur within each item. Some rephrasing of the items was required, as occasionally the second statement relied on the previous reading of the first. In each case where phrasing was changed, the alterations were minimal.

- Ford and Clarke use a numbered scale for students to indicate their preference for a particular statement. The current design used boxes for the student to mark, this avoided any unconscious weighting as to “1” being “good” and higher numbers “bad”, or vice versa.
When I'm reading a book (or other information source) for my studies, I prefer to spend quite a long time skimming over and dipping into it to get a clear picture of what it's about and how it will be relevant.

Figure 6.1: Example of an item from the SPQ.

The GRE test was constructed from items derived from a GRE test primer (Duran, Powers & Swinton, 1987) by Cox, Stenning and Oberlander (1995). It contains two kinds of item which have conventionally been termed as verbal reasoning/argument analysis problems, and model, or analytic reasoning, items. There were several problems of each type, many of which had 4 or 5 questions and students had 35 minutes to work through the test. The verbal scale (GREV) was out of a maximum 10, and the analytic reasoning scale (GREA) was out of 12.

Stenning and Cox's (1995) version of the immediate inference task was used, which, like Newstead's (1995) version offers three response categories: "yes", "no", and "can't tell". In Stenning and Cox's version, there was also a graphical section, where statements had to be determined as true, false, or independent of various Euler's Circles representations. This part of the test was omitted, because indicating ways of representing syllogism premisses may have biased students towards using these diagrams in their syllogism solution before teaching and may also have affected the learning of the methods for solving syllogisms.

In addition, a short questionnaire designed to gauge the current and past academic background of the student was devised. Students were asked about their current course of study, their previous studies, and whether they had ever studied formal logic, Venn diagrams, or Euler's Circles.

Median splits were performed on the GREA, the GREV, the PFT, and the HFT in order to distinguish groups of students who scored low and high on each test. Students were classified as "no-convertors" or "non-convertors" on the basis of their response to no-conversion.

---

1 The graphical section originated with Newstead (1989), however there was only response categories true and false in this study. Stenning and Cox (1995) also offered the opportunity for responding "can't tell" in the graphical test.
6.1.4 Teaching strategies

On the basis of results in the pre-tests, students were matched in pairs and assigned to either the EC method or the ND method for solving syllogisms. This meant that there were no main effects of any pre-test score between the two teaching groups. The pre-tests were marked anonymously, so the scores of the individual were unknown to the instructor as the syllogism method was taught.

Before being shown the syllogism solution strategy, all students were given 8 syllogisms and asked to "voice their thoughts" as they solved these. 5 syllogisms had valid conclusions, and 3 had no valid conclusion. The 5 syllogisms with valid conclusions were chosen to represent a range of difficulties according to the studies of Johnson-Laird (1983) and Ford (1995). In addition, these syllogisms were chosen to include some items that "spatial" reasoners found easier to solve than "verbal" reasoners, and some items that "verbal" reasoners found easier than "spatial" reasoners, according to Ford's criteria and data. Table 6.1 indicates the percentages of correct responses for students in each study. Ford did not study syllogisms with no valid conclusion, but a study by Bara, Bucciarelli and Johnson-Laird (1995) did study adults responses to these problems. Numbers for these rows relate to percentage of subjects correctly responding "no valid conclusion." In the instructions to the student it was stated that "some" was to be taken to mean "at least one and possibly all". Following the student's first response, but only at the student's

<table>
<thead>
<tr>
<th>syllogism</th>
<th>Johnson-Laird</th>
<th>Bara et al.</th>
<th>Ford</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Bs are As, Some Bs are Cs</td>
<td>65</td>
<td>80</td>
<td>43.7</td>
</tr>
<tr>
<td>All As are Bs, All Bs are Cs</td>
<td>95</td>
<td>86.7</td>
<td>60</td>
</tr>
<tr>
<td>All As are Bs, Some Bs are Cs</td>
<td>-</td>
<td>10</td>
<td>-</td>
</tr>
<tr>
<td>No As are Bs, Some Bs are Cs</td>
<td>5</td>
<td>40</td>
<td>13.3</td>
</tr>
<tr>
<td>All Bs are As, No Bs are Cs</td>
<td>0</td>
<td>13.3</td>
<td>18.7</td>
</tr>
<tr>
<td>No Bs are As, No Bs are Cs</td>
<td>-</td>
<td>70</td>
<td>-</td>
</tr>
<tr>
<td>Some As are Bs, Some Bs are Cs</td>
<td>-</td>
<td>15</td>
<td>-</td>
</tr>
<tr>
<td>Some As are Bs, No Bs are Cs</td>
<td>20</td>
<td>37.5</td>
<td>66.7</td>
</tr>
</tbody>
</table>

Table 6.1: Percentage of correct responses to syllogisms by study.
request, the "correct" answer to the syllogism was given by the experimenter. The pre-test syllogisms were coded for first response and final response, if this differed, and the response times in seconds were recorded.

Then students were either taught the EC method or the ND method. After the students were instructed in the method, the same 8 syllogisms used in the pre-instruction test were solved using the taught method. Not all students completed all 8 syllogisms, as the student was not allowed to proceed to the next problem until they had come to the correct answer, and the session was concluded after a one hour time limit.

Students were videoed as they worked through the teaching task, and the dialogues between student and instructor were transcribed. In a separate study (Monaghan & Stenning, 1998), speed of solution of syllogisms was not related to accuracy, so for assessing the effectiveness of acquiring a particular method these measures seem to be inappropriate. In the results section, latencies and the number solved are only reported to provide a general indication of the range of response in the group. In order to detail the students' responses to the teaching method and to partial out this effect of speed, the protocols were marked for errors and interventions. An error was recorded when the student incorrectly applied the method to the problem, and interventions were instances where the instructor offered assistance to the student. Interventions only occurred when the student had not spoken or written for some time, or when the student asked for help and seemed unable to continue unassisted.

### 6.2 Results

The results section is divided into three parts. The first part reports the pre-test measures, including the responses to the II test. The second section analyses syllogism solution prior to instruction, comparing responses to the pre-tests. The third part compares the two methods of solving syllogisms and relates them to the pre-test measures.
<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFT</td>
<td>12.82 (3.35)</td>
</tr>
<tr>
<td>EFT</td>
<td>18.19 (7.99)</td>
</tr>
<tr>
<td>GREA</td>
<td>8.10 (2.34)</td>
</tr>
<tr>
<td>GREV</td>
<td>4.62 (1.96)</td>
</tr>
</tbody>
</table>

Table 6.2: Mean and SD for pre-test measures.

<table>
<thead>
<tr>
<th>Correlations</th>
<th>HFT</th>
<th>GREA</th>
<th>GREV</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFT</td>
<td>0.45*</td>
<td>0.37†</td>
<td>0.26</td>
</tr>
<tr>
<td>HFT</td>
<td>0.03</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>GREA</td>
<td></td>
<td>0.49*</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.3: Correlations among pre-test measures.

6.2.1 Pre-tests

Table 6.2 shows means and standard deviations for the pre-test measures. Table 6.3 shows the correlations between scores on the pre-tests.

The GREA and GREV scores are roughly commensurate with those for the Stanford University students that participated in the original HP study (see Chapter 3). In the HP study, mean GREA score was 7.05 (2.72) and mean GREV score was 5.32 (1.25).

The significant correlation between the HFT and PFT is in accord with the literature (Jonassen & Grabowski, 1993). However, the lack of correlation between the GREA and the HFT is contrary to general expectations of finding a positive manifold among any “ability” measures (Carroll, 1993). This lack of correlation is perhaps due to the small sample size, but is not problematic in the context of the current study, due to the current emphasis on stylistic variation rather than measures of ability. Due to the high concentration of data, the microgenetic approach to studying problem solving does not rely so heavily on such generalisations.

For the serialist/holist questionnaire, one student did not express a preference for any learning...
style, 12 students were classified as holists, and 9 as serialists. Holists scored higher on the HFT than serialists, with mean scores of 21.55 and 13.00, respectively (t(18) = 2.83, p<0.02). Holists also seemed to score higher than serialists on the PFT, with means of 13.83 and 11.22, respectively (t(19) = 1.85, p = 0.08). Holists seemed to score higher on the GREA than serialists: 8.55 to 7.33, but the difference is not significant (t(18) = 1.15, p = 0.27). Scores for holists and serialists on the GREV scale did not differ (means of 4.45 and 4.78, respectively).

For the academic background questionnaire, all students reported little or no exposure to Venn diagrams, Euler's Circles, or logic training, therefore all students can be considered naive with regard to the syllogism method they were to learn. Academic background did not relate to any of the other pre-test measures, so it is not considered further in this study.

Eleven students responded "true" to no-conversion (hereafter, this group is termed "Convertors"), two students answered "false", and nine students responded "can't tell" (Non-convertors). The two students that responded "false" are omitted from further analyses.

No significant difference in PFT score was found between Convertors and Non-convertors (means of 12.00 and 14.00, respectively, t(18) = 1.32, p=0.21). No difference was found for HFT score, either (the mean for Convertors was 20.55, and that for Non-convertors was 16.63, t(17) = 1.10, p=0.29). In line with the predictions, Convertors scored lower on the GREA test than did Non-convertors (means of 7.45 and 9.38, respectively, t(17) = 1.97, one-tailed p<0.05).

Table 6.4 compares the serialist/holist classification with the Convertor/Non-convertor distinction. Though not significant ($\chi^2(1) = 2.77$, $p<0.1$), there is a suggestion of a result conforming with predictions: students that are classified as serialists tend to be Convertors, and students that are holists tend to be Non-convertors.

<table>
<thead>
<tr>
<th></th>
<th>serialist</th>
<th>holist</th>
</tr>
</thead>
<tbody>
<tr>
<td>convertor</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>non-convertor</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 6.4: Comparing serialist/holist preference to no-conversion.
6.2.2 Performance on the syllogisms prior to instruction: spontaneous strategy development

The 8 syllogisms given to the students prior to their instruction provides an opportunity to assess their style of reasoning prior to training. Responses to the syllogisms indicated a broad range of accuracy and speed of solution. The mean number of syllogisms solved "correctly" was 5.7 out of 8, with a maximum of 7, and minimum of 3. For syllogisms that the student responded to "correctly", the slowest student averaged 137 seconds, while the quickest student averaged 22 seconds. The mean over all subjects and problems was 76 seconds.

Besides offering an indication of the baseline performance for the student's ability to solve syllogisms, the pre-treatment data provide an opportunity to describe the spontaneous strategies developed by the students in solving syllogisms. The students' protocols were analysed for behaviour that resembled the "spatial" or the "verbal" strategies as described by Ford (1995). Unfortunately, several of the subjects could not be classified as using one of these strategies, as their descriptions of the method they used to solve the syllogisms did not provide a very compelling sense of direct access to the processes in operation. For example, one student when asked how he came to the conclusion "all As are Cs" from the syllogism "All As are Bs, All Bs are Cs" says:

it just seemed obvious to me because all the architects are botanists and all the botanists are chemists ... it's just the way they are.

Using Ford's criteria, three students can be classified as using a spatial strategy, and 7 students as using the verbal strategy. 7 students' methods remained unclassifiable. The difficulty in assessing the strategies used by the students is not surprising due to the small number of problems used in the current experiment. In Ford's study, some students do not demonstrate use of one or other of the strategies before solving several problems. One of Ford's subjects, for example, does not indicate use of the verbal strategy until explaining the method she used on the 21st problem she attempted. It may be that students take time either to develop their strategies, or to become aware of the strategy they are using.

An alternative way of deciding a student's solution strategy is by assessing whether their answer patterns correspond more closely with predictions from the spatial method or the verbal method.
According to Ford's data, the syllogisms can be ranked according to how difficult they ought to be to a student using the spatial strategy, or using the verbal strategy. Data is only available for the syllogisms with valid conclusions. Then, for each student, the syllogisms can be ranked according to the response time in getting to the correct conclusion. This analysis assumes that syllogisms that are more difficult for a particular strategy require longer processing time in order to reach a conclusion. If the correct conclusion was not produced for one or more syllogisms these syllogisms were ranked joint lowest in the student's profile. Spearman's correlation coefficient then provides a "distance" measure of the ranking of the student's responses to the syllogisms with that of the two proposed scales generated from the strategies. That Spearman's coefficient is not reliable for such small sets of data does not matter in this case, as the test is a measure of the "nearness" of the student's ranking to the predicted results from one of two strategies, therefore confidence parameters are not relevant. A student whose ranking more closely resembles that of the spatial method is judged to be using this method predominantly, even if the student does not indicate explicit awareness of this strategy use.

Assessing performance in this way, 6 students were classified as using a spatial method, and 11 students were classified as using the substitution method. This way of assessing pre-instructional strategy did not relate to that derived from the students' explanations: of the 10 joint classifications, only 4 were in agreement (binomial test p>0.7). The classifications derived from the Spearman's correlation coefficient were then measured against the psychometric scores from the pre-tests. No comparisons reached significance, though 5 of the 7 serialists fitted the verbal profile, whereas only 4 of the 9 holists fitted this profile ($\chi^2 = 1.17$, p=0.28). No relationship was found between no-conversion and the spontaneous strategies ($\chi^2 = 0.00$, p>0.9). The pre-teaching syllogisms do not seem to reveal much in the way of different strategies in reasoning.

6.2.3 Learning the methods for solving syllogisms

The EC group

Generally, the group learned the task well, solving a mean of 6.89 syllogisms in the time allowed. The minimum number of syllogisms solved using this method by a subject was 4, and four subjects managed to solve all 8 syllogisms. Response times varied from a mean minimum of 66 seconds to 285 seconds, with the average time across all students for all syllogisms being 154
seconds.

Number of errors in applying the method varied greatly for the students. Mean number of errors was 7.9, with minimum of 1 and maximum of 20. There was no correlation found either with the number of problems solved with the method or the mean response time of using the method to solve a problem. As anticipated, this measure was not found to relate to the pace that the student works at. There is, however, a negative correlation between the number of syllogisms correctly solved in the pre-instructional syllogisms and the number of errors made in the course of applying the method ($r = -0.70$, $p<0.05$). This means that the student’s pre-instructional ability seems to have some effect on the ease with which the new method is applied.

The ND group

Students were generally slower at using the ND method: the mean number of syllogisms solved using this method was 3.88, with the range varying from 3 to a maximum of 5 syllogisms solved. Response times varied from a mean minimum of 169 seconds for each question to a maximum of 525 seconds, with an average of 351 seconds for all students for all problems for the ND method.

Comparing the two groups

Comparing the spontaneous strategies used by students to their performance on the teaching methods did not reveal much of interest. Those that were classified as using Ford’s “verbal” strategy solve more syllogisms with the ND method than do Ford’s “spatial” strategy students (means of 4.5 syllogisms and 3.25 syllogisms, respectively, $t(7) = 2.18$, $p<0.02$). When errors made in using the method are the dependent variable, the result approaches significance$^4$: “verbal” students made 3.75 errors on average, whereas “spatial” students made 5.5 errors ($t(6) = 2.18$, $p=0.07$). However, no interaction was found between teaching method and pre-instructional strategy, as “verbal” students seemed to make fewer errors on the EC method as well (though this was not significant).

The pre-test groups were compared to the teaching methods by using 2-way unrelated ANOVAs, with teaching method and pre-test grouping as factors, and the number of errors or interventions as the dependent variable. The means for each grouping are shown in Table 6.5.

$^4$This is significant as a one-tailed t-test ($p<0.05$).
Errors Interventions

<table>
<thead>
<tr>
<th></th>
<th>PFT-Lo</th>
<th>PFT-Hi</th>
<th>PFT-Lo</th>
<th>PFT-Hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>12.25</td>
<td>4.40</td>
<td>33.25</td>
<td>13.40</td>
</tr>
<tr>
<td>ND</td>
<td>5.33</td>
<td>4.20</td>
<td>31.33</td>
<td>33.60</td>
</tr>
<tr>
<td>serialist</td>
<td>holist</td>
<td>serialist</td>
<td>holist</td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>11.60</td>
<td>2.00</td>
<td>30.40</td>
<td>7.00</td>
</tr>
<tr>
<td>ND</td>
<td>4.75</td>
<td>4.50</td>
<td>31.00</td>
<td>34.50</td>
</tr>
<tr>
<td>GREA-Lo</td>
<td>GREA-Hi</td>
<td>GREA-Lo</td>
<td>GREA-Hi</td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>12.00</td>
<td>4.60</td>
<td>29.50</td>
<td>16.40</td>
</tr>
<tr>
<td>ND</td>
<td>4.20</td>
<td>5.33</td>
<td>28.40</td>
<td>40.00</td>
</tr>
<tr>
<td>Convertor</td>
<td>Non-convertor</td>
<td>Convertor</td>
<td>Non-convertor</td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>9.33</td>
<td>2.50</td>
<td>24.67</td>
<td>9.00</td>
</tr>
<tr>
<td>ND</td>
<td>3.67</td>
<td>5.20</td>
<td>25.67</td>
<td>37.00</td>
</tr>
</tbody>
</table>

Table 6.5: Comparing EC and ND methods by pre-test groupings.

Overall, the PFT-Hi group made fewer errors and required fewer interventions than the PFT-Lo group. For errors, there was a main effect of PFT group ($F(1,13) = 6.87, p<0.05$), and an interaction approaching significance between PFT group and teaching method ($F(1,13) = 3.84, p<0.1$). For interventions, there were no main effects, but there was a significant interaction between PFT group and teaching method ($F(1,13) = 6.33, p<0.05$). These interactions are shown in Figure 6.2. PFT-Hi students make fewer errors and need fewer interventions than PFT-Lo students on the EC method (errors: $t(7) = 2.58, p<0.05$; interventions: $t(7) = 3.20, p<0.02$). PFT-Hi students require more interventions on the ND method than the EC method ($t(8) = 3.67, p<0.01$). PFT-Lo students seem to make more errors on the EC method than the ND method ($t(5) = 2.05, p<0.1$).

For serialist-holist group, for errors, there was a main effect of serialist-holist group ($F(1,12^5) = 9.32, p<0.05$), and an interaction between serialist-holist group and teaching method ($F(1,12) = 8.39, p<0.02$). For interventions, there are main effects of serialist-holist group ($F(1,12) = 5.54, p<0.05$) and of teaching method ($F(1,12) = 11.05, p<0.01$), and an interaction between group and teaching method ($F(1,12) = 10.12, p<0.01$). Figure 6.3 indicates the interactions. Serialists make

---

5The one student that did not demonstrate a preference for one learning style was omitted from the analyses.
Figure 6.2: Number of errors and interventions by PFT group for the teaching methods: $F(1, 13) = 3.84, p<0.1$, and $F(1, 13) = 6.22, p<0.05$, respectively.

more errors and require more interventions than holists on the EC method (for errors: $t(6) = 3.11, p<0.05$; for interventions: $t(6) = 3.82, p<0.01$). Holists seem to make fewer errors on the EC method than the ND method ($t(5) = 2.21, p=0.08$) and require many fewer interventions for the EC method than the ND method ($t(5) = 5.08, p<0.005$). Conversely, serialists make fewer errors on the ND method than the EC method ($t(7) = 2.58, p<0.05$), though the difference in interventions for the methods is not distinguished.

Figure 6.3: Number of errors and interventions by serialist/holist group for the teaching methods: $F(1, 12) = 8.39, p<0.02$, and $F(1, 12) = 10.12, p<0.01$, respectively.
Figure 6.2: Number of errors and interventions by PFT group for the teaching methods: \( F(1, 13) = 3.84, p<0.1 \), and \( F(1, 13) = 6.22, p<0.05 \), respectively.

More errors and require more interventions than holists on the EC method (for errors: \( t(6) = 3.11, p<0.05 \); for interventions: \( t(6) = 3.82, p<0.01 \)). Holists seem to make fewer errors on the EC method than the ND method (\( t(5) = 2.21, p=0.08 \)) and require many fewer interventions for the EC method than the ND method (\( t(5) = 5.08, p<0.005 \)). Conversely, serialists make fewer errors on the ND method than the EC method (\( t(7) = 2.58, p<0.05 \)), though the difference in interventions for the methods is not distinguished.

Figure 6.3: Number of errors and interventions by serialist/holist group for the teaching methods: \( F(1, 12) = 8.39, p<0.02 \), and \( F(1, 12) = 10.12, p<0.01 \), respectively.
For GREA group, for errors, there were no main effects but an interaction of GREA group and teaching method (F(1,13) = 5.64, p<0.05). For interventions, there was a main effect of teaching method (F(1,13) = 5.26, p<0.05), and an interaction of GREA group and teaching method (F(1,13) = 6.33, p<0.05). The interactions are shown in Figure 6.4. GREA-Hi students seem to make fewer errors than GREA-Lo students on the EC method (t(7) = 2.31, p=0.05). For the ND method, GREA-Hi students require more interventions than GREA-Lo students (t(6) = 2.91, p<0.05). GREA-Hi students require fewer interventions on the EC method than the ND method (t(6) = 2.80, p<0.05), whereas GREA-Lo students make fewer errors on the ND method (t(7) = 2.95, p<0.05).

![Figure 6.4: Number of errors and interventions by GREA group for the teaching methods: F(1, 13) = 5.64, p<0.05, and F(1, 13) = 5.26, p<0.05, respectively.](image)

For no-conversion, there was an interaction with teaching group that approaches significance (F(1, 12) = 3.25, p=0.10). For interventions, there is a significant interaction with teaching group (F(1, 12) = 6.38, p<0.05). Convertors perform equally on both teaching methods, but Non-convertors require very few interventions on the EC method and many more on the ND method (see Figure 6.5).

### 6.3 Discussion

The correlation between the PFT and the GREA test and the relation between each of these tests and the measure of serialist/holist learning style were as hypothesised. This supports the per-
spective taken of these tests as reflecting similar styles of processing underlying representational differences. The serialist/holist measure requires the student to state how they believe they approach learning situations, and somewhat surprisingly, they seem to be able to do this with some degree of accuracy. The GREV test was not found to relate to any measure other than the GREA, due in part at least to the fact that both subscales of the GRE occurred on the same test. What this test is assessing is not illuminated by the correlations. This is true also for the HFT, which relates only to the PFT.

In accord with the predictions, no-conversion related to the psychometric measures. Non-convertors score higher on the GREA test than Convertors, and seem also to be those students with a holistic preference for learning. This points towards the usefulness of the II as a stylistic indicator.

For spontaneous strategies, gauging strategies in Ford’s (1995) terms was found to be difficult from such a small sample of problem answers. The results were not very fruitful. This highlights the advantage of using “enforced” strategies for examining the response of students in a short time-scale. Nonetheless, there was a link between using Ford’s spatial strategy and easily acquiring the EC method, which supports claims of their similarity.

From the graphs of the interactions, it can be seen that number of errors and number of interventions seem to measure slightly different aspects of performance, though they are correlated (across both EC and ND groups, r(15) = 0.51, p<0.05). Number of errors is taken to be a measure of the safety of a method – a high score means the student is more likely to go wrong in the

Figure 6.5: Number of interventions by No-conversion by teaching method (p<0.05).
method. The number of interventions is taken to reflect the students' ease of applying the method—how much direction they require in following the method correctly. Either measure can be used to assess the ease with which the student acquired the method.

A successful outcome of the teaching study was that the two theoretically motivated taught methods seemed to be comparable. There were very few main effects of teaching group for number of errors and interventions in the ANOVAs, but there were several interactions with the pre-test groups. The relationship between acquiring the EC method and prior use of Ford's spatial strategy also supports claims for the analysis of learning strategies as reflecting the propensities and preferences of students.

The specific hypotheses raised with regard to teaching syllogism solution will be discussed first. The extent to which the more general aims of the study, as presented in Chapter 5, were achieved will then be considered.

The GREA was predicted to relate differently to the two teaching methods, with GREA-Hi students acquiring the EC method more easily and the GREA-Lo students finding the ND method more in keeping with their reasoning style. GREA-Hi students made few errors on both methods, and required fewer interventions for the EC method than the ND method. GREA-Lo students were found to make more errors on the EC method than the ND method, but were not distinguished in terms of number of errors made on each method. The predictions regarding the relationship between the GREA measure and the teaching methods were generated by the original HP studies, where GREA-Hi students were those that prefer fewer, abstract cases, and GREA-Lo students prefer to generate several specified cases. This distinction seems to carry into syllogistic reasoning also.

There were no specific predictions made with regard to the GREV subscale, and the teaching syllogisms experiment has done little to illuminate this measure. This lack of relationship between the measure and response to the teaching methods is encouraging for one reason, however, indicating that it is a very different measure to the GREA test, which highlights robust differences in learning the two methods.

It was predicted that serialist students would find the ND method easier, and that holist students would find the EC method easier. Even though the ND method was not distinguished according to this grouping, it was still the case that holists found the EC method easier than the ND method, and serialists found the EC method harder to acquire than the ND method. This lends
further support to the connection between the serialist/holist style and preference for the different implementations of the ICIA in syllogism solution. The serialist/holist distinction is founded on preferences for different learning strategies, and is not formulated in terms of representations. Yet a strong connection emerges between the style measure and use of methods invoking different representations.

The HFT did not distinguish response to the two teaching methods, this is perhaps due to the high level of structure in the two methods. The HFT is a measure of differences in the way students build up the structure of the stimulus. In the teaching methods this structure is already present.

If “spatial ability” reflects flexibility in using different strategies for solving problems, then PFT-Hi students were predicted to acquire both methods more easily than PFT-Lo students. There was a main effect found of PFT group on error, but this difference seems to be mainly due to performance on the EC method, where PFT-Hi students make fewer errors and require fewer interventions. PFT-Hi students’ greater ability and flexibility seems to be only for graphical material. This suggests that the PFT may be a style rather than an ability measure. The extent to which the PFT measures flexibility in translating between representations that differ in terms of the level of abstraction they employ has not been addressed by this study. This will be pursued in the HP study in the next chapter. If spatial ability is this ability to translate, then it may reflect on the EC method and not on the ND method due to the EC method using different levels of abstraction in representation to the ND method. Conventionally, it has been assumed that graphical representations are more specific, and in the comparison between the EC method and the ND method it was shown that the ND method draws on a more expressive system. So, the EC method demonstrates a larger difference in level of abstraction employed in the representations. An analysis relating the psychometrics to different stages in using the two solving syllogism methods will contribute towards this debate.

As predicted, Non-convertors made fewer errors and required fewer interventions on the EC method, but required more on the ND method. However, Convertors did not find the ND method easier, nor did they find the EC method harder. Non-convertors more easily acquire a method where the symmetry of the “no” premiss is represented, but find a method that does not respect this symmetry (the ND method) harder. This seems to suggest that students that do not initially accept the symmetry are more comfortable with a method that indicates this symmetry. This
may be because the graphics make explicit cases that the student was unsure held previously (see Bauer & Johnson-Laird, 1993). This is not inconsistent with the view that Non-convertors have an affinity with parallel/abstract approaches to problem solving – the critical individual still must be specified, but only after parallel resolution of all individuals. Convertors, however, were not found to respond differently to one or other method. Thus, some support is gained for no-conversion relating to the sequential or parallel implementation distinction. Non-convertors are those that respond better to representations that implement the parallel/abstract algorithm.

In terms of the general aims of the study, there were several successes. The aptitude-treatment interaction in the HP experiment between GREA group and teaching method for transfer was replicated in the syllogistic domain. Here, in a shorter intervention the same effect is found for acquiring a method as was found in HP for transfer of skills from a method. In both syllogistic teaching and HP it proves to be important to fit the method to the style of the student as gauged by the GREA test. The success of the replication supports the contention that the abstract characterisation of the different strategies that are used by students underly both reasoning domains. Stenning and Monaghan (in press) argue that ATIs are seldom found to replicate because tasks and the processes brought to bear on those tasks are not well understood. The replication here is due to locating the relevant similarities between the domains: similarities in terms of strategies, representations and algorithms that characterise the distinct approaches.

The GREA can therefore be seen as a linchpin connecting accounts in terms of strategy, representation and algorithms that underly performance on reasoning tasks in the syllogistic and HP domains. For a full explanatory account, a more detailed investigation into the GRE test is required. The GREA seems to be very much a measure of style rather than ability – the GREA-Lo students are good at learning from certain presentations, and the GREA-Hi students are poor at learning from the same ND method. This was a second main aim of the study: to distinguish issues of style from ability, by relating the psychometrics to processing styles. For the other measures, too, this approach meets with some success. No-conversion was found to be a useful diagnostic tool for reflecting strategic differences, as was the serialist/holist dimension. The PFT groups were found only to relate differently to the EC method, so some style dimension is relevant here. However, severing the link between the PFT and ability is by no means complete, and the approach of the next section is to refine the link rather than replace it. The study has not contributed to understanding the relationship between strategy and the HFT or the GREV measures.
First, though, the psychometric measures can be given a more detailed profile by assessing the different stages in using the two methods. This may help to tease apart the different measures and assist in the subsidiary aim of providing computational accounts of the psychometrics, as well as offering a more detailed investigation into the two teaching methods. These finer-grained distinctions in performance for different stages in the task are presented for consideration in tandem with the general interactions observed across the whole method.

As mentioned in the last chapter, the two methods can be directly compared for three stages: translating into the representation, manipulating the representations, and translating out of the representation to make a conclusion. Distinguishing these stages and relating errors and interventions for each stage to the psychometric groupings enables a distinction between the measures for reflecting translating between representations and manipulating representations. For the translating-in stage, if the pre-tests interact with teaching method then this suggests that the representation is an important determiner in the strategy, and then the representation can be seen to promote and precede certain strategic approaches. The differences in the two implementations of the ICIA centre around the manipulating stage, so the pre-tests are expected to distinguish between the taught methods in this stage insofar as they reflect these alternative approaches. Most of the general interaction effects are anticipated to be due to differences on the manipulating stage. As with the translating-in stage, the translating-out stage will reflect the extent to which the representation reflects the strategy that it invokes.

Errors and interventions for each of the three stages were measured, and analyses for each stage are presented in the next section.

### 6.4 Comparing different stages in the two methods

Two-way unrelated ANOVAs were performed, with teaching method and pre-test group as factors and the number of errors or interventions for each stage in using the methods as dependent variable. Only significant interactions of teaching method and pre-test group are reported below. As with the general comparison between responses to the two methods, both the GREV test and the HFT groups did not indicate differences in learning from the two methods.
6.4.1 Translating-in stage

On the translation-in stage, there were few errors made or interventions required, as this stage in the procedures can be accomplished by reading off the translations. One interaction was, however, significant for this stage: GREA group by taught method on the number of interventions required (Figure 6.6): $F(1, 13) = 5.30, p<0.05$, with no main effects. For the EC method, GREA-Hi students require fewer interventions than GREA-Lo students ($t(7) = 2.92, p<0.05$). GREA-Hi students require fewer interventions on the EC method than for the ND method ($t(7) = 3.82, p<0.01$).

6.4.2 Manipulating stage

For the manipulating representations stage, both serialist/holist group and GREA group interacted with taught method for both dependent variables. No-conversion also interacted with taught method for interventions, and approached significance for an interaction with taught method for errors.

For the serialist/holist group, the interaction with taught method for number of errors on this stage was significant (Figure 6.7): $F(1, 12) = 5.89, p<0.05$, with a main effect only of learning style ($F(1, 12) = 7.12, p<0.05$). For number of interventions, $F(1, 12) = 5.46, p<0.05$, with a main effect only of taught method ($F(1, 12) = 15.05, p<0.01$). Serialists make more errors and require more interventions than holists on the EC method (for errors: $t(6) = 3.04, p<0.05$; for interventions: $t(6)$...
Figure 6.7: Number of errors and interventions for manipulation stage by serialist/holist group for the teaching methods: $F(1, 12) = 5.89, p<0.05$, and $F(1, 12) = 5.46, p<0.05$, respectively.

Holists require fewer interventions on the EC method than they do on the ND method ($t(5) = 5.20, p<0.005$), whereas serialists seem to make more errors on the EC method than they do on the ND method ($t(7) = 1.93, p<0.1$).

Figure 6.8: Number of errors and interventions for manipulation stage by GREA group for the teaching methods: $F(1, 13) = 7.45, p<0.02$, and $F(1, 13) = 5.50, p<0.05$, respectively.

GREA group and taught method interacted both for number of errors (Figure 6.8) ($F(1, 13) = 7.45, p<0.02$, with no main effects) and interventions ($F(1, 13) = 5.50, p<0.05$, with a main effect only.
of taught method: $F(1, 13) = 11.84, p<0.01)$. GREA-Hi students make fewer errors on the EC method than do GREA-Lo students ($t(7) = 2.43, p<0.05$). For the ND method, GREA-Lo students seem to require fewer interventions ($t(6) = 1.95, p<0.1$). GREA-Hi students make fewer errors ($t(7) = 2.54, p<0.05$) and require fewer interventions on the EC method than they do on the ND method ($t(6) = 3.53, p<0.02$).

![Figure 6.9: Number of interventions for manipulating stage by no-conversion group for the teaching methods: $F(1,12) = 6.80, p<0.05$.](image)

For the no-conversion groupings, number of errors by taught method approached significance ($F(1, 12) = 4.20, p=0.06$), and there were no significant main effects. Convertors make fewer errors on the ND method than do Non-convertors ($t(6) = 2.83, p<0.05$). For interventions there was an interaction between no-conversion group and taught method ($F(1, 12) = 6.80, p<0.05$), and a main effect only of taught method ($F(1, 12) = 13.76, p<0.005$). Non-convertors seem to require fewer interventions on the EC method and more on the ND method than do Convertors (Figure 6.9). For the ND method, Non-convertors require more interventions than Convertors ($t(6) = 2.37$, one-tailed $p<0.05$), and Non-convertors require more interventions on the ND method than the EC method ($t(5) = 5.29, p<0.005$).

### 6.4.3 Translating-out stage

For the translating-out stages, both serialist/holist style and PFT group interacted with taught method.
Figure 6.10: Number of errors and interventions for translating-out stage by serialist/holist group for the teaching methods: \( F(1, 12) = 5.23, p<0.05 \), and \( F(1, 12) = 9.82, p<0.01 \), respectively.

For the serialist/holist groups, number of errors were significant (Figure 6.10): \( F(1, 12) = 5.23, p<0.05 \), with a main effect only of learning style \( F(1, 12) = 5.23, p<0.05 \). Also for number of interventions, \( F(1, 12) = 9.82, p<0.01 \), with a main effect again only of learning style \( F(1, 12) = 5.89, p<0.05 \). Holists make fewer errors and require fewer interventions than serialists on the EC method (for errors: \( t(6) = 2.59, p<0.05 \); for interventions: \( t(6) = 3.21, p<0.02 \). Serialists make fewer errors on the ND method than they do on the EC method \( (t(7) = 2.52, p<0.05) \). Holists, on the other hand, require more interventions on the ND method than they do on the EC method \( (t(5) = 3.27, p<0.05) \).

For PFT group there was an interaction with taught method both for number of errors (Figure 6.11): \( F(1, 13) = 4.90, p<0.05 \), with a main effect of taught method \( F(1, 13) = 9.96, p<0.01 \) and PFT score \( F(1, 13) = 9.31, p<0.01 \) and number of interventions: \( F(1, 13) = 7.16, p<0.02 \), with a main effect only of PFT score: \( F(1, 13) = 6.66, p<0.05 \). PFT-Hi students make fewer errors and require fewer interventions than PFT-Lo students on the EC method (for errors: \( t(7) = 3.31, p<0.02 \); for interventions: \( t(7) = 3.33, p<0.02 \). PFT-Lo students make more errors and require more interventions on the EC method than they do on the ND method (for errors: \( t(5) = 3.26, p<0.05 \); for interventions: \( t(5) = 3.08, p<0.05 \).
6.5 General discussion

Individuating different stages in using the two methods indicate points where the psychometric measures diverge, and this contributes to the aim of aligning psychometric with computational investigations of individual differences in reasoning. Each of the four measures that revealed individual differences over use of the whole method related to different stages in using the two methods.

Only one measure – GREA – related to the translating-in stage. Here, GREA-Hi students find translating into the EC method easier than translating into the ND method, and GREA-Lo students seem to find the EC harder and the ND translations easier. This suggests that GREA grouping reflects the sensitivity of students to representations that promote certain strategic approaches. The representations themselves are semi-independent of the algorithms that underly them (the algorithm can, at least, be implemented in one of several ways), yet still ease of use is stylistically determined by the same dimension that determines differences in acquiring use of the strategies.

Most differences between the taught methods were predicted to be evident in the manipulating stages, as this stage in using the two methods implements one or other of the two strategies. These expectations were supported by the results. The serialist/holist groups did not seem to
distinguish performance on the ND method. However, taking response to the ND method as a benchmark indicates that serialist/holist group does distinguish response to the EC method. Serialists make more errors on the EC method than on the ND method, and holists make fewer errors and require fewer interventions on the EC method than on the ND method. This suggests that response to the parallel/abstract approach is strongly predicted by serialist/holist group. Serialists find parallel/abstract approaches difficult to acquire and utilise, whereas holist students find this approach fits well with their style.

GREA grouping also distinguishes performance on the manipulating stage, and performance differs on both teaching methods. For the EC method, GREA-Lo students find it more difficult than GREA-Hi students. For the ND method, the reverse is found: GREA-Lo students find this method easier than do GREA-Hi students. This pattern is confirmed by both errors and intervention measures. Though the serialist/holist dimension does not distinguish both teaching methods, the GREA groupings do.

No-conversion is the third measure that relates to the manipulating stage in using the two methods. Non-convertors required fewer interventions on the EC method than the ND method, and Convertors required fewer interventions than Non-convertors on the ND method for this stage. This measure relating to the manipulating stage may be particularly surprising, as the grouping is determined by response to interpreting premisses, so it might be thought to reflect only translating-stages. This adds further support to the claim that what is required for a full account of the phenomena of individual differences is a consideration of strategic, representational and algorithmic variation.

PFT grouping was only found to relate to the translating-out stage in using the two methods. PFT-Lo students make more errors and require more interventions on the EC method than the ND method. They also make more errors and need more interventions on the EC method than PFT-Hi students. There is no evident difference in response on the ND method, so the PFT groupings can be seen to relate to translating from the graphical to sentential representations. Translating between sentential representations does not distinguish performance. Perhaps it is this ability to translate that relates to ability to switch strategies within a task, as the availability of representations is greater.

For the translating-out stage, the serialist/holist grouping distinguishes performance on both methods. Serialist students make more errors than holists in translating from the EC method.
Serialists also make more errors on the EC method than they do on the ND method. Holists and serialists do not make a different number of errors or require different numbers of interventions on the ND method, but holists do require more interventions on the ND method than they do on the EC method. Serialists find the ND method more conducive to their style, whereas holists seem to find the EC representations easier to translate into a conclusion.

The general interactions between pre-test grouping and the taught methods can be seen to be largely accounted for by interactions for particular stages in using the two methods. The GREA-effect is due to interactions with method for translating-in and manipulating stages. The serialist/holist dimension determines differences for manipulating and translating-out stages. No-conversion determines differences in the manipulating stage. Finally, the PFT distinguishes performance on the translating-out stage.

The absence of significant results with the HFT and the GREV groupings mean that no insight into these measures has been achieved by linking psychometrics to a reasoning task. As previously mentioned, the HFT may be a better predictor when the task is less structured and freedom for strategy selection is given. The GREV remains a mysterious measure, but the contrast with strong effects for the GREA scale suggest that it is a very different kind of problem to the analytic reasoning problems. This distinction will be explored further in the next chapter.

The results of this study have indicated that reasoning is stylistic: interactions between teaching methods and psychometric groups together with the absence of main effects show that certain groups respond favourably to different methods. Different representations have been characterised as relying on alternative processes, and these relate to cognitive styles. Constructing representations and translating from representations – stages in the procedure that might be expected to be independent of process if differences can be defined as purely representational – have also been shown to be stylistic, in that certain students seem to acquire one method more easily than the other for each stage in using the methods. The results also indicate that translating-in and translating-out of representations are very different processes. GREA group relates to the former only, PFT group was found only to relate to the latter. Studies of using external representation to support reasoning must not assume skills of translating are symmetrical. A consequence of this study is to indicate that use of external representations must be distinguished into (at least) these three stages in order to assess and identify their use for supporting reasoning (Scaife & Rogers, 1996; Cox & Brna, 1995).
How does the current study relate to Ford’s (1995) study on the development of spontaneous strategies for syllogistic reasoning? Unfortunately, due to the small number of syllogisms used in the pre-instructional test, it was difficult to categorise students as using either Ford’s verbal or spatial strategy. A rough attempt to make such a categorisation resulted in a paucity of results, with the only significant relation being between using a verbal strategy and more easily acquiring the ND method. Thus, connections to Ford’s study remain, on the whole, theoretically motivated.

It is hypothesised that the development of Ford’s verbal strategy is due to a preference for a sequential/specific approach, and developing Ford’s spatial strategy is based on a preference for a parallel/abstract approach. This contrasts with accounts that presume the differences are in terms of preferences for representing in a particular modality. This may be the case apparently, but the differences must be described also in terms of strategies and algorithms. Describing the two strategies in terms of the differences in the implementation of the same class of algorithms highlights the ties between representations and strategies.

What does the current study contribute to understanding strategic variation and development in other complex reasoning tasks? The three studies reviewed in Chapter 2 found that strategies developed to express an appropriate level of abstraction for the task. These tasks demonstrated variation in the effectiveness of acquiring different strategies. In contrast, the methods for solving syllogisms taught to students seem to be equally effective for solving the task, though this is modulated by cognitive style. The other reasoning tasks were initially presented to justify the generality of strategic variation in terms of abstract/concrete properties of the representations used. One principle design of this study was that both strategies should be equally appropriate for solving the problems, and thus differences between students should be in terms of the processes that underly the apparent representational differences. The styles that have been shown to be important predictors of learning from each method are not just to do with the level of abstraction in the representation, but also address the different processes elicited by an approach that either specifies properties of individuals or leaves them abstract. The current study shows the plausibility of this interpretation, particularly when the links to the HP study discussed in Chapter 3 are considered. There, the relationship between level of abstraction and strategy is explicit, and the overlap with the syllogistic study is manifested in terms of the GREA test, and the serialist/holist distinction.

As yet, nothing has been added to the investigations of strategy development, but this important
and illuminating issue will be addressed in the next chapter. The only source of a direct empirical link between the teaching syllogisms study and the reported studies on strategies in other reasoning domains is the measure of "spatial ability" used in each. Subjects that scored high on spatial ability tests were those that were good at developing and assessing strategies that invoked an appropriate level of abstraction. In the teaching syllogisms study, PFT-Hi students had less difficulty in translating out of both types of representation to form a conclusion. Stenning and Monaghan (in press) suggest that this is due to the PFT reflecting the ability to compare verbal "folding narratives" and spatial patterns (see Snow, 1980). It does not appear to be a general ability to use representations that employ different levels of abstraction, but it may be the ability to translate from a specific into a more expressive system. In the syllogism study, students that use the EC method must translate from the EC representations into "natural language"; the ND method requires translations from a more expressive system than the EC diagrams – a fragment of first order logic. In the other reasoning domains discussed in Chapter 2, spatial ability relates to the use of non-spatial representations, which can be seen as the translation of information from a specific, spatial representation system to more expressive systems of representation. Further investigations into the PFT as a strategic indicator are discussed in the next chapter.

The replication of the HP aptitude-treatment interaction adds credence to the generality of the observed stylistic behaviour. It also means that a more general account is required so that similarities across domains can be established and accounted for. The replication rests on the assumption that learning transferable skills from the HP courses relates to easily acquiring taught methods for solving syllogisms. This equivalence is supported by comparing the different algorithms underlying the EC/ND methods with the holist/serialist proof styles in HP. But the GREA test remains a black box in the account. It seems to predict stylistic preference for sequential/specific versus parallel/abstract approaches, but the test itself has not been analysed in the current study. One aim of the next chapter is to unpack the GRE test, both GREA and GREV items, and attempt to understand why it relates to stylistic variation in terms of use of abstract/concrete representations. It is clear, though, that it is not a measure of flexibility, as certain strategies are more easily acquired by GREA-Lo students. Similarly, in the Stenning, Cox and Oberlander (1995) study of transfer of problem solving skills from logic courses GREA-Lo students benefitted from the non-graphical version of HP, whereas GREA-Hi students got worse at certain sorts of problem after following the course. When representations that promote parallel/abstract approaches are avail-
able (EC method/HP), GREA-Hi students seem to excel. When representations that channel sequential/specific approaches are only available (ND method/non-graphical HP), these students perform less well than GREA-Lo students.

6.6 Holes in the scholastic sieve

There are two main issues arising from the syllogism teaching study. One is the necessity of an account of the GRE test, in order to determine the nature of the overlap between HP and syllogisms. The results thus far have indicated that the GRE is a useful test of strategic preference, but what is it about the problems it presents that provides this strong indication of cognitive style?

Second, an explicit investigation into the relationship between abstract/concrete representation use and different computational styles would be a useful extension of the syllogism study. Relating psychometrics to HP would provide such a link. Furthermore, this would provide an account of strategy development, relating the use of enforced strategies in the syllogism experiment to those developed spontaneously. A study that assesses the way students learn will also help to finesse the hypotheses about what and why certain students learn from certain presentations of information. The transfer of skills from following the HP course is underdescribed as yet. Such an assessment of strategy change would also assist in providing a more detailed account of what spatial ability tests are really testing. Why do certain students change the level of abstraction in their problem solving strategies, and how does this relate to the sequential/specific parallel/abstract distinction?

HP is chosen as the domain for further study because it is fruitful for spontaneous strategies, and also enables detailed assessment of the level of abstraction used in representations. Also, its shared perspective with the teaching syllogisms experiment allows for some generalisation of the results. The next chapter presents a further investigation into problem solving in HP, relating performance to various psychometric measures. In particular, the GRE is considered in terms of the appropriate levels of abstraction to be used to effectively solve each problem. This occupies the first part of the chapter. The second part of the chapter reports the experiment as a replication and extension of the initial investigations into transfer of problem solving skills from following a logic course.
Chapter 7

A continuum of problem types and developing strategies: Hyperproof II

7.1 A taxonomy of problem types

The motivation for providing a taxonomy of problem types is to distinguish reasoning problems in terms of the level of abstraction required in representations appropriate for solving them. Hence, problems solved effectively using concrete representations will be distinguished from those that are better solved using representations with greater abstraction.

This taxonomy will have two applications:

- It will distinguish the problem types in the GRE test, and ability to solve these different problems will emerge as the difference between students that score high and low on the two subscales of the GRE. Scores on this test relate to differences in the selection of representations of varying levels of abstraction.

- It will provide a way of distinguishing the different types of goal in HP, and enable an assessment of proof strategies in terms of choosing varying levels of abstraction in solving problems. This will provide a finer-grained insight into the mechanisms of strategy change as a result of following the HP logic course, and provide a means of assessing the use of these strategies for different problem types in the tests of transfer of reasoning skills. In this way, the initial explorations of Chapter 3 will be extended.
In addition, the classification of problem types indicates the varying appropriateness of strategies that employ a sequential-specific or parallel/abstract approach. Problems that are better solved using a more abstract level of representation may actually be better solved using a parallel/abstract approach, whereas those best solved with concrete representations are almost certainly better solved using a sequential-specific approach.

7.1.1 Problem types in the GRE

The GRE is traditionally divided into two subscales: the GREA and the GREV. To reiterate, GREA problems are those that are often solved by using diagrams (Cox & Brna, 1995) and have explicit constraints on the number of models consistent with the given information. GREV problems are infrequently solved by using diagrams, and the variety and number of models is difficult to see from the problem, as they seem to be more about comparing argument structure than constructing models in order to solve the problem. This immediately suggests that problems on the GRE test vary in terms of the level of abstraction required in the representations used to solve these problems. This level of abstraction is the basis of the classification of problems.

Consider Figure 7.1 (reproduced from Figure 3.3). This is one problem from the GREA subscale. The information in the problem constrains a single model, so fully concrete representations are useful for solving this problem\(^1\). This problem will be called a single-case problem, and can be found both in the BW test and in HE. For the GRE test, it is referred to as a GRES problem. This problem is appropriately solved by constructing the fully constrained model and then reasoning with this concrete representation.

An alternative approach to solving this problem is to assess the potential arrangements of offices and people, and then testing these arrangements for consistency with the given information. For this problem, there are 6 people to be arranged in 6 offices, offering 120 permutations. This number of cases is going to make such a strategy very inappropriate. One linguistics student when presented with this problem in a discussion class on the uses of representation in problem solving started to panic, and said she just couldn't begin to solve this sort of problem. This would certainly be the case if an attempt to operate on all possible arrangements was undertaken.

Consider now the GRE problem shown in Figure 7.2. This problem is consistent with many

\(^1\)Though every student may not use external representations to solve this problem (Cox, 1996).
An office manager must assign offices to six staff members. The available offices, numbered 1-6 consecutively, are arranged in a row, and are separated only by 6-foot-high dividers. Therefore, voices, sounds, and cigarette smoke readily pass from each office to those on either side.

Miss Braun’s work requires her to speak on the telephone frequently throughout the day.

Mr White and Mr Black often talk to one another in their work, and prefer to have adjacent offices.

Miss Green the senior employee, is entitled to Office 5, which has the largest window.

Mr Parker needs silence in the office(s) adjacent to his own.

Mr Allen, Mr Parker and Mr White all smoke.

Miss Green is allergic to tobacco smoke and must have non-smokers in the office(s) adjacent to her own.

1. The best location for Mr White is in Office 1, 2, 3, 4 or 6?

2. The best employee to occupy the office furthest from Mr Black would be Mr Allen; Miss Braun; Miss Green; Mr Parker or Mr White?

3. The three employees who smoke should be placed in Offices 1,2 and 3; 1, 2 and 4; 1, 2 and 6; 2, 3 and 4; or 2, 3 and 6?

Figure 7.1: Analytic reasoning item in the GRE test.

different situations, however students often construct graphical representations in order to solve this problem. There are 6 poets, and students can assume one of two attitudes to each poet. This means that there are $2^6 = 64$ different possibilities for students’ poet-liking profile. The different universal rules in the problem limit the number of possible profiles. The first clause outlaws 16 situations, as does the second clause. The fourth clause outlaws a further sixteen situations, 2 of which coincide with situations outlawed by the first clause, and a further 2 coincide with situations outlawed by the second clause. The seventh clause reduces the possible situations by 16, of which 4 have previously been ruled out by clause 1 and four by clause 2. This means that after the constraints have been in place, there are 28 outlawed situations leaving 36 possible situations. This is far more than the GRES problem, but it is further reduced by the statement in the question. “Enjoying the poetry of Browning” reduces the number of situations by 32, of which 16 have already been ruled out. This means that to solve the problem presented requires the assessment of 20 situations, so a more appropriate representation for this problem would
Professor Kittredge's literature seminar includes students with varied tastes in poetry. All those in the seminar who enjoy the poetry of Browning also enjoy the poetry of Eliot. Those who enjoy the poetry of Eliot despise the poetry of Coleridge. Some of those who enjoy the poetry of Eliot also enjoy the poetry of Auden. All those who enjoy the poetry of Coleridge also enjoy the poetry of Donne. Some of those who enjoy the poetry of Donne also enjoy the poetry of Eliot. Some of those who enjoy the poetry of Auden despise the poetry of Coleridge. All those who enjoy the poetry of Donne also enjoy the poetry of Frost. Mister Huxtable enjoys the poetry of Browning. He may also enjoy any of the following poets except

A. Auden
B. Coleridge
C. Donne
D. Eliot
E. Frost

Figure 7.2: Multiple case problem in the GRE test.

employ a greater degree of abstraction than that required for the GRES problem. As several situations are consistent with the given information, this is called a "multiple-case" problem, in the GRE context: a GREM problem.

An alternative approach could avoid using abstraction in the representations by assessing each of the 64 different possible poet-liking/disliking combinations. Though there are not as many cases as with the office allocation problem, this would still be extremely inefficient as an approach.

Figure 7.3 (reproduced from Figure 3.4) indicates a problem originally from the GREV subscale. This problem does not present immediately specifiable constraints on the number of models consistent with the given information. One possibility is that there are three binary properties in the problem shown in Figure 7.3: safe/dangerous, industrial Hg/other sources of Hg; and Hg in water/other sources. Together these define $2^3 = 8$ possible situations, and the statements determine which of these situations turn out to be the case. This means that a strategy that assessed each possibility would be much more plausible for this type of problem than for the office allocation
Excessive amounts of mercury in drinking water, associated with certain types of industrial pollution, have been shown to cause Hobson's Disease. Island R has an economy based entirely on subsistence-level agriculture; modern industry of any kind is unknown. The inhabitants of Island R have an unusually high incidence of Hobson's Disease.

Which of the following can be validly inferred from the above statements?

- I Mercury in drinking water is actually perfectly safe.
- II Mercury in drinking water must have sources other than industrial pollution.
- III Hobson's Disease must have causes other than mercury in drinking water.

II only; III only; I or III but not both; II or III but not both; or II or III or both?

Figure 7.3: Verbal reasoning item in the GRE test.

or poets problem. Attempting to construct a representation that respected the constraints of the problem would, however, be rather difficult for this problem, as the cases are extremely difficult to identify, and the information would have to be disembedded from the problem. A greater degree of abstraction would be appropriate for this problem. Perhaps the linguistics student was attempting to go through each case exhaustively regardless of the structure of the problem, which would explain her panic at the office allocation problem, and the relative ease with which she solved the mercury poisoning problem.

Problems on the GRE test present a dual continuum depending on how the complexity of the problem is assessed. If the size of the problem space is considered then the office allocation problem will be harder than the poet's problem, which in turn is harder than the mercury problem. If the number of situations consistent with the constraints are considered then the GRE problems present a continuum running in the other direction: ranging from single case through multiple case problems, which vary over the number of cases consistent with the given information, to the most verbal that offer only a few, difficult to determine constraints on the possible situations. This classification system ignores other ways that the problem types differ. For example, the extent to which content is buried in form, or the extent to which induction or deduction plays a role in problem solution may also be contributory factors. But these other ways of describing different problem types in the GRE may well be concomitant properties of the different representational systems.
A hierarchical cluster analysis of responses to the various problems in the GRE test reflects intuitions about the different problem types. The clustered responses from the 21 students that took the GRE test in the teaching syllogism (see Chapter 6) study are shown in Figure 7.4.

Clustering shows for which problems students tended to answer the same way. This means that problems that are closely clustered together meant the same students did well on both, or poorly on both. "Office" is the GRES problem shown in Figure 7.1. "Poets" is the GREM problem shown in Figure 7.2. Note that these two problems are closely clustered. The "interviews" problem was initially in the verbal subscale of the GRE test, but it does provide several constraints on the number of situations that are consistent with the given information and many students use external representations to assist with their problem solving for this question. "Verbal set 1" refers to three questions, including the verbal problem in Figure 7.3 as well as two other problems that seem similar in requirements to this question. "Gubernatorial" is a single case problem, and "verbal set 2" is a second set of verbal problems. These last two problem groups occur at the end of the test, so responses to these are closely clustered primarily because most students failed to complete any of these problems. For the other four questions, no relation between position in the test and response emerges.

The clustering suggests the following ordering of problems in terms of the number of situations they determine consistent with the given information, from fewest on the left to most on the right:

office ←→ poets ←→ interviews ←→ verbal set 1

The alternative conception of the problems as ranged along a continuum according to the size of
the problem space they are presented in offers the same ordering, from largest problem space on the left to smallest on the right.

This supports the assumption that solving problems that require abstract representations and solving problems that require more concrete representations is prone to stylistic variation: students respond similarly to "analytic" items, but differently to "verbal" items. There is also support for conceptions of the dual-scale of the GRE test (Stenning, Cox & Oberlander, 1995; Yang & Johnson-Laird, 1999), in terms of the expressivity and specificity of representational systems. However, a further subdivision of the "constraint-satisfaction" problems distinguishes the GRE into three scales: single-case (GRES), multiple-case (GREM), and verbal (GREV) problems.

There is some empirical support for the taxonomy of problem types in terms of the level of abstraction required for solving the problem. However, for the system of classification to be useful it has to be shown to be general, and to reflect performance in other problem solving tasks. The next section indicates that the distinction can be extended to different problems in the BW test.

7.1.2 Generalising the taxonomy: problem types in the BW

A BW test problem is shown in Figure 7.5. A diagram is given and abstraction is present in the situation in terms of labels being unspecified for the blocks. There are verbal statements that present restrictions on the way labels are assigned to the objects in the diagram.

As with the GRE test, there are problems that fully specify the situation that must be assessed in order to answer the question – these are the single case problems (BWS). There are also problems that require several situations to be assessed, and these are multiple case problems (BWM). These are likened to the GREM problems rather than the GREV problems as the given graphical situation presents immediate constraints on the possible situations consistent with the given information. There is not a corresponding problem type in the BW to the GREV problems.

Two additional types of problem are distinguishable in the BW test. The first of these are questions about logical independence rather than logical consequence. An example is Statement B in Figure 7.5. The situation where "b is a medium dodecahedron" is not determined by the given information, as a state of affairs where b is one of the small dodecahedrons is also consistent with the given information. These are called non-consequence problems (BWN).

The final type of problem in the BW test is a class of questions about possible situations consistent
Suppose you are given the following information about the blocks shown above.

I. Block b is a dodecahedron and block a is closer to the front than b.
II. Block a is not small.
III. Block d is closer to the left side of the grid than a.
IV. If block d is a tetrahedron, then it is not small.

Based on the above information, which of the following statements must be true?
A. Block a is a large tetrahedron.
B. Block b is a medium dodecahedron.
C. Block d is a small cube.
D. Both A and B.
E. Both A and C.
F. A, B, and C.

Figure 7.5: Example of a BW problem.

with the given information. They are generally phrased in the form: “which of the following statements might be true”. These are classified differently to the other goal types, and are termed BWmodal problems.

So, in the BW test, the following problem types are distinguished: singe-case (BWS), multiple-case (BWM), non-consequence (BWN), and the “might follow” problems (BWmodal).

Justification for the taxonomy spanning different tasks will arise if the same students find, for example, single case problems easy in both tests. However, there are differences between the GRE and the BW and these generate hypotheses about potential differences in use of abstraction in problem solving. The first is a difference in the surface features of the test, the second is more fundamental to different strategies based on constructing representations using abstraction to varying degrees.

One difference between the BW and the GRE test is that problems in BW have multiple choice responses, and this means that each problem may simultaneously involve solving different problem types. This is a surmountable obstacle, however, and the solution provides a rich data set.
Consider the BW problem in Figure 7.5. The correct answer is (A). Statement (A) requires only one situation to be constructed to prove it, so it is classified as a BWS problem. Statements (B) and (C) do not follow from the given information, so these are classified as BWN problems. This means that responses from (A) through to (F) score differently on the BWS and BWN scale. Table 7.1 shows the scores for each response.

<table>
<thead>
<tr>
<th>Response</th>
<th>BWS score</th>
<th>BWN score</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 7.1: Scoring system for the BW problem.

The more basic difference between the GRE and the BW is that the BW presents a graphical situation to the student which is then augmented by applying sentential information to the problem. For the GRE test, no such situation is given, and if a graphical representation is used by the student then it has to be constructed from scratch. Cox (1996) has identified this “constructivist” difference in the use of external representations by students: use of a given situation and generation of a situation are different skills.

7.1.3 Problem types and strategies in HP

The different problem types in the GRE and the BW test are also identifiable as different goal types in HP. For each problem in HP, there may be several goals that have to be achieved. The requirements of these goals vary in terms of whether it is a sentential expression or a graphical situation that has to be assessed, and whether proofs of consequence or non-consequence are required. Of the proofs of consequence, it may be that a single situation is determined by the given information, or that several situations that have to be constructed are required. These different goal types are considered in more detail below, along with a discussion of the range of strategies available to solve them.
Given information: 
\[ \forall v \, (\text{Small}(v) \rightarrow \exists y \, (\text{Small}(y) \land \text{Cube}(y) \land \text{Adjoins}(v, y))) \]
\[ \forall v \, (\text{Medium}(v) \rightarrow \exists y \, (\text{Medium}(y) \land \text{Cube}(y) \land \text{Adjoins}(v, y))) \]
\[ \sim \exists v \, \text{Large}(v) \]

Can you determine the shape of the highlighted block?

Can you determine the shape of the highlighted block?

Can you determine the shape of the highlighted block?

Can you determine the shape of the highlighted block?

Can the following sentence be a consequence of the given information?
\[ \exists x \exists y \,(\text{Cube}(x) \land \text{Cube}(y) \land x \neq y) \]

Figure 7.6: The Hyperproof interface – Problem 4.

Figure 7.6 (Figure 3.5 shows different strategies used to solve this same problem) is a HP problem with four goals indicated on the right of the Figure.

Goal 4 requires the student to assess the validity of a given *sentential* statement: \( \exists x \exists y \,(\text{Cube}(x) \land \text{Cube}(y) \land x \neq y) \). These are termed S-goals. The method for achieving S-goals varies according to whether the sentence has to be shown to be a consequence or a non-consequence of the given information. In each instance, however, the S-goal is achieved either by using natural deduction rules in HP, or by using rules that support the translation of information from graphical situations to a sentential expression. This goal type is distinguished from those that require the assessment of a graphical situation as translating is a key feature, potentially placing different requirements on the level of abstraction appropriate to achieving the goal.

Goals 1 to 3 require the student to assess graphical situations: whether the shape of each of the three objects can be determined. Goals 1 and 3 are about situations that are *not* determined by the conjunction of graphical and sentential information, i.e., they are proofs of *non-consequence* (GN-goals). These are akin to BWN problems. One situation consistent with the given information is where the leftmost block is a small tetrahedron and the other two blocks are small cubes. Another consistent situation is where the rightmost block is a small tetrahedron and the other two blocks
are small cubes. Constructing these two situations and justifying their consistency with the given information by invoking a rule called “check the assumptions” is sufficient for achieving goals 1 and 3 in HP. Alternatively, students may prove each of goals 1 and 3 separately, constructing two situations for each goal. Varying use of situations to solve these problems does not necessarily indicate different use of graphical abstraction, but rather illustrates whether goals are achieved in parallel or successively.

Goal 2 requires a proof of consequence. This class of problem is called a GC-goal. In goal 2 in question 4 this is done by exhausting the possible situations consistent with the given information and then indicating their commonality: that the middle block is a cube. Variability in the number of situations used reflects the “granularity” of the representations covering the space of possible situations. An alternative method for achieving this goal is to create situations which show the three possible shapes of the middle block and indicate that two of the possible shapes mean a contradiction with the given information. This results in a shorter proof, and one that involves a greater degree of abstraction in the situations, as more information is left unexpressed. Another method for solving this problem is to translate the critical information into sentential expressions and use the natural deduction rules of HP. This results in an even shorter proof, using representations that are more abstract. This goal is one type of the GC-goals, one which requires splitting into multiple cases (GCM-goal). This proof by splitting into cases is a feature of HP promoted by its designers, who see case-based, heterogeneous reasoning as a natural way of approaching problems (Barwise and Etchemendy, 1992). Proofs by constructing graphical cases is an alternative and complementary strategy to using proofs composed of sentential rules, for example.

Another type of GC-goal is one that can be achieved by the application of sentential information to a situation in a single step (GCS-goal). An example of this type of goal is shown in Figure 7.7. In this question, the GCS-goal is Goal 1, as the information in the problem is sufficient to determine the labelling of the leftmost block as “c”. Students can use different strategies for solving these problems. One way of solving the problem is by invoking a rule called “apply” that justifies the application of sentential information to a graphical situation. Applying the information of the 2nd, 3rd and 4th sentences means that this can be achieved in one step, but labels for all objects have to be applied (so the rightmost block has to be labelled “a” in the same situation). An alternative method is to apply labels one at a time. Several students construct a situation
supported by an “apply” rule that names the rightmost block “a” invoking the 1st, 3rd and 4th sentences. They then construct another situation that adds the label to the leftmost block using “apply” supported by the 3rd and 4th sentences. Objects remaining unlabelled means that the abstraction of the situations is greater, so students that achieve the goal using one situation (i.e., one use of “apply”) utilise more concrete external representations to support their reasoning. Students that use staged “apply” rules use more abstract representations.

**Figure 7.7: GCS goal in HP.**

Another strategy used to solve the GCS-goal in Figure 7.7 is to exhaust the two different ways that labels can be applied, and to indicate that one application of labels is inconsistent with the given information, and the other application of labels is consistent. This is a property of the particular structure of this problem, so for most other GCS-goals the comparison is between an “abstract” strategy where the required situation is built up gradually, and a “concrete” strategy where the situation is achieved in a single step.

To summarise, GCS-goals are problems where the information constrains a single situation sufficient for achieving the goal. GCM-goals are problems where the information does not constrain a single situation: there are several situations consistent with the given information, each of which must be constructed in order to achieve the goal. These are comparable to the single-case and
multiple case problems as described in the GRE and the BW test. HP offers a window into the use of level of abstraction in representation, and this is reflected in the number of situations used by the student in order to achieve the goal. For GCS-goals, few situations in a proof reflects use of less abstraction in the graphical representations, and use of more situations reflects use of more abstract representations. For GCM-goals, this is reversed, with few situations reflecting more abstraction in the situations used, and more situations indicating that each situation is more concrete.

7.2 Aims and expectations

The syllogism study left the GREV subscale a mystery. It did little better in terms of illuminating the GREA subscale, but indicated that the GREA aptitude-treatment interaction was robust across different reasoning domains. The analysis of the GRE test in this chapter relates the different question types to use of different levels of abstraction, and advocates a triple partition of the measure. The HP study presented in this study will address these issues directly, by distinguishing different problem types in the GRE, as well as in the BW task, and relating them to strategic performance in HP. This not only extends the investigation of the psychometric measures, but also broadens the study to involve spontaneous strategy use and development as well as use of given strategies.

7.3 Method

84 Philosophy students at the University of Gothenberg participated in the experiment, taking a logic course as part of their studies. Students at this university come from a broader academic background than generally found at Edinburgh University or Stanford University (the previous subject populations). There are also fewer academic prerequisites for courses. The pre-tests and post-tests were taken voluntarily by students.

Students followed the HP course using Barwise and Etchemendy’s (1994) materials as part of a logic course, which also exposed students to a more traditional natural deduction method of logical proof (Bennet, Haglund, Westerståhl & Sönströd, 1997, derived from Mates, 1965).
At the end of the course, students were set six HP problems. Each problem was designed to be "indeterminate", so that the graphical situations given in each problem were graphically abstract to varying degrees. Each problem contained several goals, so that several tokens of each goal-type as discussed above were given to the students\(^2\). Students were free to solve problems in their own time. The proofs were computer-logged.

7.3.1 Pre-tests and post-tests

Pre-tests and post-tests were chosen in an attempt to directly replicate the original HP study, and also to extend it to include psychometric measures for comparison with proof styles.

As in the original study, a version of the BW test was given pre- and post-course. The BW test was marked for overall score as with the original study, as well as marked in order to assess response to the different problem types within the test. The GRE was also administered pre- and post-course, and was marked for scores on the GREA and the GREV subscales as well as for each problem type. All versions of the BW and the GRE tests were identical to those used in the original study. Students had 30 minutes for the GRE test and 15 minutes for the BW test.

In the pre-test students were also given the PFT and the HFT (French, Ekstrom & Price, 1963), as used in the teaching syllogisms study. As previously mentioned, the SPQ for testing serialist/holist learning style was not used as the experiment was begun before the SPQ emerged as an interesting gauge of variation in problem solving approach. All pre- and post-tests were presented in English. The course coordinator felt the students' grasp of English was sufficient for this not to be an issue in the results.

7.4 Results

61 students did the HFT and the PFT, 57 students did the pre-course BW test, and 59 students did the GRE test; 72 students completed the HP exam problems; and 27 students did the BW and GRE post-course test. Some of the HP records were lost due to bugs in the logging program, so n-values vary throughout the analyses. Full data for all questions and pre-tests exists for 39 students. Full data for all questions, pre-tests and post-tests exists for only 20 students.

\(^2\)Grateful thanks to Cecilia Sönströd for collecting the data, and to Jon Oberlander for creating the HP exam problems.
The results are divided into three sections. The first examines the overall scores from the BW and the GRE tests, to explore the extent to which the original HP study is replicated with this different population. The second section distinguishes scores on the different problem types in the BW and the GRE and examines the relationship between these scores and the psychometric measures. The third section relates different strategies used in HP to the psychometrics and to the transfer of reasoning skills to the BW and the GRE tests.

### 7.4.1 Replicating the original HP study

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFT</td>
<td>12.66 (3.28)</td>
</tr>
<tr>
<td>HFT</td>
<td>14.02 (7.28)</td>
</tr>
<tr>
<td>GREA</td>
<td>4.86 (2.57)</td>
</tr>
<tr>
<td>GREV</td>
<td>2.71 (1.87)</td>
</tr>
</tbody>
</table>

Table 7.2: Mean (SD) for pre-tests.

<table>
<thead>
<tr>
<th></th>
<th>GREV</th>
<th>HFT</th>
<th>PFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>GREA</td>
<td>0.30* (57)</td>
<td>0.34* (55)</td>
<td>0.27* (55)</td>
</tr>
<tr>
<td>GREV</td>
<td></td>
<td>0.22 (55)</td>
<td>0.22 (55)</td>
</tr>
<tr>
<td>HFT</td>
<td></td>
<td></td>
<td>0.26* (59)</td>
</tr>
</tbody>
</table>

Table 7.3: Correlations between pre-tests.

Table 7.2 shows mean and standard deviation for the group. GREA and GREV scores are lower than for the Stanford and Edinburgh students, whereas PFT and HFT scores are more closely comparable. Table 7.3 indicates the correlations among the pre-test measures of GREA and GREV subscale scores, and the HFT and the PFT. The correlations are similar to those found in the teaching syllogisms study, which lends some authority to comparing student groups across these different subject populations.

As with the original study on learning from HP, students were divided into those that scored higher and lower on the GREA test according to a median split. The pre- and post-test scores on
the BW as a whole were then compared for these groups. As with the original study, the GREA-Hi students improved their score on the BW more than did the GREA-Lo group (t(23) = 1.80, one-tailed p<0.05). This interaction is shown in Figure 7.8. Again, the robustness of the aptitude-treatment interaction effect supports comparison between the different experiments, paving the way for an extension and more detailed study of the intricacies of strategy development and transfer of reasoning skills.

Another result from the original HP study was that GREA-Hi and GREA-Lo students were distinguished by the style of the proofs that they constructed. A primary feature of these different proof styles was that GREA-Hi students employed more abstraction in the graphical situations they used, i.e., they used fewer fully concrete situations than GREA-Lo students. Does this replicate in the current study?

The GREA-Hi students used fully concrete situations 40% of the time, whereas GREA-Lo students used them slightly less at 37% of the situations they constructed. However, the HFT grouping reflected this stylistic difference better. HFT-Hi students use concrete situations only 32% of the time, whereas HFT-Lo students' proofs are considerably more concrete, using 51% fully concrete situations, which is significantly more (t(51) = 2.29, p<0.05).

The proof styles in the current study do not replicate those found in the original HP study. There are two potential explanations for this. One possibility is that the GREA subscale is testing something slightly different in the Swedish student group. As they are taking the test in their second language, an additional component of "verbal" ability may have been in place. Another possible explanation is that the effects of differences in teaching (in particular, the inclusion of a natural deduction method in teaching) may have influenced learning. However, there are still
differential effects of GREA-Hi/GREA-Lo groupings in terms of how well skills are transferred to the BW test, so the former explanation seems less likely. This suggests that the latter is the key explanation, and so more detailed examination of the range of strategies used in HP is required.

Examining the proofs of one particular question, one that was also used in the original HP study and that produced a broad range of problem solving strategies (shown in Figure 7.6) indicated a different range of strategies to those found in the original HP study. In the original study, proofs varied in terms of the use they made of graphical abstraction. GREA-Hi students used proofs that employed a greater degree of abstraction, building up concrete situations in stages by applying sentential information gradually. GREA-Lo students used concrete situations more readily. In the current study, the concrete graphical strategy was very much in evidence, but the abstract graphical situation was much rarer than in the original study (6 out of 69, compared to the original 9 from 22). In addition, a third method for solving the problems was used by 11 students in the new study. This employed the sentential rules of HP, meaning that very little use of graphical situations was made. This will be called the “sentential” strategy. The sentential HP strategy is like an extreme version of the abstract graphical strategy, as it employs a more abstract form of representation to achieve the goals. Empirical justification for this continuum may be drawn from scores on the pre-test measures, as these are taken to represent different attitudes to the use of representations employing different levels of abstraction. Table 7.4 shows pre-test scores for each strategy used for the problem shown in Figure 7.6, that is used in both HP studies.

<table>
<thead>
<tr>
<th>Pre-test</th>
<th>Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sentential</td>
</tr>
<tr>
<td></td>
<td>graphical</td>
</tr>
<tr>
<td>GREA</td>
<td>5.00</td>
</tr>
<tr>
<td>GREV</td>
<td>2.00</td>
</tr>
<tr>
<td>HFT</td>
<td>19.10</td>
</tr>
<tr>
<td>PFT</td>
<td>13.50</td>
</tr>
<tr>
<td>BW improvement</td>
<td>1.50</td>
</tr>
</tbody>
</table>

Table 7.4: Pre-test scores for students classified by use of the three strategies for the problem occurring in both HP studies.
This continuum of strategy types is reflected in the mean scores on the HFT and the PFT: those that use most abstraction in their strategy seem to score higher on both these measures. The sentential strategy students also seem to score higher on the GREA, but the continuum is not respected for this measure. Of interest is the lack of match between scoring well on the GREV subscale and using the sentential strategy. Those that do well on GREV problems tend to use concrete graphical strategies for their HP proof.

The differences in HFT score between the sentential and the concrete graphical strategy groups are significantly different (t(44) = 2.65, p<0.02). The group including both sentential and abstract graphical strategies scored higher on the HFT than the concrete graphical strategy group (t(49) = 2.64, p<0.02). However, the HFT score for the abstract graphical group alone was not found to differ significantly from either other strategy group, partly due to the small number of students that use this strategy (n = 5).

The results indicate a partial replication of the original HP study. The general transfer of skills to the BW according to GREA group are the same, but the strategies are variable. Monaghan et al. (1999) argue that in order to justify and explain aptitude-treatment interactions, it is necessary to attend to the exact task requirements. In addition, the use of the psychometric pre-tests lends credence to the claim that these psychometrics reflect attitude to using different levels of abstraction in representation. The use of a different type of strategy in the current study for solving HP problems can be seen as a more extreme version of the abstract graphical strategy within this framework. The next section provides more detailed results pertinent to the aim of distinguishing the skills learned by students as they follow the HP course.

7.4.2 Performance on the BW and GRE problem types

For the BW test, all problem types in the pre-test are significantly correlated (see Table 7.5). In the post-test, BWS problems are significantly correlated with all other problem types, but BWM, BWN and BWmodal problems do not intercorrelate. Significant correlations between the pre- and the post-tests were found for BWS problems, and for BWM problems, but were not found for BWN problems or for BWmodal problems.

PFT group distinguished scores on several of the BW problem types. Repeated measures ANOVAs were used with problem type score as dependent variable and time as within subjects
Do students with certain psychometric profiles change score more on one BW problem type than another? Is change in the same direction for all problem types? To answer these broader questions, a 2 by 4 repeated measures ANOVA was performed, with problem type nested under time as within subjects factors, and psychometric groups as between subjects factor.
HFT predicted general difference in scores across all BW problem types and tests (F(1, 21) = 9.95, p<0.01), with HFT-Hi students scoring higher than HFT-Lo students. Similarly, PFT-Hi students scored higher overall (F(1, 21) = 13.66, p<0.005). A general effect of time approached significance (F(1, 21) = 3.87, p=0.06), but no general interactions between time and HFT or PFT group were found. An effect of problem type was found (F(1, 21) = 131.45, p<0.001) with students scoring highest on BWS problems, and lowest on BWM problems. No interactions between goal type and HFT/PFT group across time were found.

When time and goal type were considered, a significant interaction was found (F(1, 21) = 10.26, p<0.001). As a general result of following the Hyperproof course, students improve their performance greatly on BWM problems, improve slightly on BWS and BWmodal problems, but seem to get worse on BWN problems. This shows that the increase that students indicate on the BW test as a result of following the HP course is due only to certain types of problem: in particular, the BWM problems. These are problems that require a certain level of graphical abstraction to be employed in their solution for them to be solved optimally, and HP assists in teaching the use of this graphical abstraction. Presumably, this is only for certain students (the “DetHi” students in the original study), and this is reflected in the interactions between time and psychometric group for the different problem types. Although there were no interactions between time, problem type and psychometric group, when the problem types were distinguished such interactions were observed. The HFT-Hi group improved on BWM problems, but the HFT-Lo group did not, and the PFT-Hi group improved on BWN problems, but the PFT-Lo group did not. Equating the HFT-Hi/PFT-Hi groups with the GREA-Hi group, an equation that depends on conceiving of each measure reflecting development of appropriately abstract representations, indicates that only certain students develop skills in HP that assist on BWM problems which are not detrimental to BWN problem solution.

Table 7.6 summarises the changes in scores according to psychometric group. The GRE test is also open to such an investigation into the development of different types of problem solving skill as a result of following the HP course.

Correlations for pre- and post-test problem types in the GRE test are shown in Table 7.7 (as with the BW data, n=51 for pre-tests, and n=25 for post-tests). Unlike for the BW problem type data, no significant correlations were found between pre-test and post-test problem types, nor between any of the pre-test and post-test scores. This suggests that GRE problem types seem to distinguish
responses more than BW problem types. The BW test’s provision of a graphical situation may lead to more homogeneity among responses to different problem types in the BW.

As with the BW problem type scores, GRE problem types were analysed using ANOVAs with problem type score as dependent variable, time as within subjects factor, and psychometric groups as between subjects factors.

For GRES problems, PFT-Hi students score higher than PFT-Lo students (F(1, 20) = 14.28, p<0.005). Time also predicted differences, with post-test scores lower than pre-test scores (F(1, 20) = 10.12, p<0.01).

For GREV problems, there was a significant interaction between time and PFT group (F(1, 20) = 5.48, p<0.05). PFT-Lo subjects increased their score (from 0.70 to 0.90), whereas PFT-Hi students’ scores decreased slightly (from 1.07 to 1.00).

Analysing the data with a 2 by 4 repeated measures design with problem type nested under time, the psychometric groups predicted different changes for each problem type from pre- to post-test. There was an effect of PFT group over all problem types (F(2, 40) = 9.54, p<0.01): PFT-Hi students
score higher than PFT-Lo students. There was also an effect of time ($F(2, 40) = 5.29, p<0.05$), the post-test was generally found to be more difficult than the pre-test. An effect of problem type was found ($F(2, 40) = 9.85, p<0.001$), with scores on GREM problems higher than those on GRES problems, which are higher than those on GREV problems.

An interaction between PFT group and problem type approached significance ($F(2, 40) = 2.79, p=0.07$), reflecting the large difference in summed (pre- and post-test) scores for GRES problems (PFT-Lo: 3.20, PFT-Hi: 5.72), slight difference for GREM problems (PFT-Lo: 5.20, PFT-Hi: 6.57) and smaller difference for GREV problems (PFT-Lo: 1.60, PFT-Hi: 2.07). An interaction between time and problem type was found ($F(2, 40) = 4.65, p<0.02$). There is a large decrease in GRES scores (2.65 to 2.15), a similar decrease for GREM scores (3.59 to 2.92), and a slight increase for GREV scores (0.92 to 0.96).

An interaction between time, problem type and HFT group was also found ($F(2, 40) = 4.93, p<0.02$). Figure 7.9 shows the change in score for each HFT group by problem type.

HFT-Hi students decrease their score on GRES problems dramatically, register a slight increase in score on GREM problems, and improve their score on GREV problems most. HFT-Lo students increase their GRES problem score slightly, but they also get worse at GREM and GREV problems. The interaction between GRE problem type and time indicates that the skills used to solve different types of problem are differentially affected by the HP course. As with the BW results, transfer occurs between domains, but not always in a beneficial manner. In general, following the HP course is good for promoting the solving of GREV problems, but the course also seems
to interfere with successful solution of GRES problems. The results also support interpreting multiple-case problems as being somehow “in-between” verbal and single-case problems in the GRE test: for GREM problems no effect on their solution was observed as a result of following the HP course.

There seem to be interesting and important differences between the BW and the GRE problem types, the correlations between them are shown in Table 7.8. Multiple-case problem scores were correlated for BW and GRE pre-tests, but just fail to reach significance in the post-test (p=0.07). GREM pre-test score also correlates with BWS and BWN pre-test scores, and correlates negatively with BWmodal post-test. GRES pre-test score correlates with BWS post-test score, and also with BWN post-test score. GREV pre-test score correlates negatively with BWmodal post-test score but with no other variable. Performance on multiple-case problems seems to be consistent across reasoning domains, but single-case scores do not demonstrate such a clear connection.

These similarities and differences will be returned to in the discussion section. The results of analyses of strategies for the different problem types in HP are now reported.

<table>
<thead>
<tr>
<th></th>
<th>GRES pre</th>
<th>GREM pre</th>
<th>GREV pre</th>
<th>GRES post</th>
<th>GREM post</th>
<th>GREV post</th>
</tr>
</thead>
<tbody>
<tr>
<td>BWS pre</td>
<td>-0.04</td>
<td>0.30*</td>
<td>0.14</td>
<td>0.37</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>BWM pre</td>
<td>-0.03</td>
<td>0.39**</td>
<td>-0.06</td>
<td>0.03</td>
<td>0.30</td>
<td>-0.31</td>
</tr>
<tr>
<td>BWN pre</td>
<td>0.11</td>
<td>0.38**</td>
<td>0.10</td>
<td>0.12</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td>BWmo pre</td>
<td>-0.08</td>
<td>0.26</td>
<td>0.13</td>
<td>0.32</td>
<td>0.33</td>
<td>-0.05</td>
</tr>
<tr>
<td>BWS post</td>
<td>0.43*</td>
<td>-0.13</td>
<td>-0.32</td>
<td>0.05</td>
<td>0.22</td>
<td>0.12</td>
</tr>
<tr>
<td>BWM post</td>
<td>0.33</td>
<td>0.23</td>
<td>-0.22</td>
<td>-0.09</td>
<td>0.36</td>
<td>-0.11</td>
</tr>
<tr>
<td>BWN post</td>
<td>0.56**</td>
<td>0.00</td>
<td>0.01</td>
<td>0.07</td>
<td>0.27</td>
<td>0.02</td>
</tr>
<tr>
<td>BWmo post</td>
<td>0.10</td>
<td>-0.40*</td>
<td>-0.39†</td>
<td>-0.03</td>
<td>0.19</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Table 7.8: Correlations between GRE and BW problem types.

### 7.4.3 HP strategies and transfer of reasoning skills

Overall proof length did not predict learning from HP (with one exception) – the distinction into different goal types in HP is necessary for highlighting the skills learned from following the course. There was no correlation between overall proof length and the pre-test measures.
For GREA, GREV, HFT and PFT, all \( p > 0.4 \). Students were divided into those that produced longer proofs and those that produced shorter proofs overall by a median split. A repeated measures ANOVA with GREA score as dependent variable with time as within subjects factor and long/short proof length as between subjects factor was performed. Time by proof length interacted significantly \( (F(1, 22) = 8.12, p < 0.01) \), with a main effect of time \( (F(1, 22) = 7.46, p < 0.02) \). Students that produced long proofs decreased their GREA score from pre- to post-test (7.23 to 4.56), whereas students that produced shorter proofs did not decrease (6.44 to 6.82). When HFT and PFT were also entered as between subjects factors, the interaction did not hold. Repeated measures ANOVAs with BW as dependent variable did not indicate any interactions or main effects of proof length.

In order to assess interactions between using different strategies for each type of HP goal, median splits were performed on the number of situations used to solve each type of goal. So, there is a GCS-short proof and a GCS-long proof group, a GCM-short and GCM-long group, and so on. The next sections report analyses of changing scores in the GRE and the BW tests measured by the different HP strategy groups: comparing strategies on sentential HP goals to graphical HP goals; comparing consequence goals and non-consequence goals; and finally comparing single-case with multiple case goals in HP.

**S-goal and G-goal strategies**

No students were in the G-long and S-short strategy group, and only two students were in the G-short S-long group, so these groups are omitted from the following analyses. So, comparisons are between students that either used shorter strategies overall, or longer strategies overall. Figure 7.10 shows the performance of the S-short G-short strategy group and those that were classified as being in the S-long G-long group on the GRE problem types.

The S-short G-short students seem to improve their scores on the GREV problems from pre- to post-course, but their scores reduce on the GRES problems. This pattern is reversed for the S-long G-long group. A repeated measures ANOVA with single/verbal problem type nested under time interacted significantly with these strategy groups \( (F(1, 17) = 5.36, p < 0.05) \). There were no significant interactions when GREM problem scores were included in the ANOVA. GREM pre-test score did, however, relate to the strategy groups: the S-short G-short group scored lower initially on GREM problems \( (t(14) = 2.13, p = 0.05) \).
Developing strategies that improve scores on the GREV problems seems to have a detrimental effect on scores in the GRES problems. Strategies that maintain or improve GRES scores seem to impair performance on GREV problems. Long strategies tend to use less abstraction in the representations, and, as anticipated, this is good for the single case goals in the GRE test, but damaging to the solution of GREV problems as these require a greater degree of abstraction than that practiced by these students. Using short strategies seems to be an effective approach for solving GREV problems, but not good for GRES problem solution. Students seem in general to develop strategies that are either abstract or concrete and these are only appropriate for solving one type of GRE problem. Furthermore, these strategies seem to have been developed to counteract weaknesses in representation use before the HP course was taken.

Figure 7.10: Change of scores for GRE problem types by HP S-goal and G-goal strategy groups.

Figure 7.11: Change of scores for BW problem types by HP S-goal and G-goal strategy groups.
Comparing performance on BW goal types for these groupings, the effect of single case problems is reversed (see Figure 7.11). The S-short G-short group increased their score on BWS problems, whereas S-long G-long students decreased their score for these problems (the change in BWS problem score was significant: t(18) = 2.57, p<0.02). Similarly, for BWN and BWmodal problems, the S-short G-short group increased their scores, whereas the S-long G-long group decreased their performance on these problems (for differences in score, t(18) = 1.88, p=0.08 for BWN problems, and t(18) = 3.26, p<0.005 for BWmodal problems). Unlike with the GRE problem types, there is no interaction between different types of BW problem and the strategy groups which means there is no indication that the different strategies developed in HP affect BW problem types differently. In general, following the HP course seems to be good for all students for solving BW problems regardless of their strategy choice.

Short proofs for a given goal type should mean that students use a level of abstraction appropriate to that problem. This is supported by the BW test results where S-short G-short students improve in general, whereas S-long G-long students do not show a general improvement, only increasing scores on BWM problems. S-short G-short students improve their GREV scores, whereas S-long G-long students do not, meaning that the predictions are supported for abstract problems. But generally shorter solutions in HP do not relate to improving GRES problem score, in fact the reverse holds. Further discriminations of G-goal types are required to cash out this puzzling finding.

The current analysis also offers clues to the psychometric profiles of students that develop different types of strategy in HP. Students that develop short strategies for both S- and G-goals in HP score lower on GREM, BWS and BWmodal pre-test problems. There are no significant differences between these two strategy groups in terms of PFT or HFT score. Students that develop shorter strategies for HP goals are initially less able at solving problems that require more abstract representations, and seem to develop these more abstract strategies as a response to this initial weakness in their reasoning.

GC-goal and GN-goal strategies

There are no non-consequence goals in the GRE test, so direct comparison with the different GN-goal strategies used in HP are not available. But the possibility still remains that distinctions in terms of GN- and GC-goal strategies may relate to different GRE problem types. Figure 7.12
shows mean scores for GRE problem types by these strategy groups.

Both the GC-long GN-short and the GC-long GN-long groups did not seem to improve their performance on any of the GRE problem types. The GC-short GN-short and the GC-short GN-long groups both demonstrated a decrease in GRES score and an increase in GREV score from pre- to post-test. This suggests that different strategies for GC-goals alone reflects differences in learning to solve the various types of GRE problems. A repeated measures ANOVA with GRES/GREV problem type nested under time interacted significantly with GC strategy group. GC-short strategy students increase their score on GREV problems and decrease their score on GRES problems. The GC-short group end up scoring higher on GREV problems (one-tailed t(20) = 1.85, p<0.05) having improved their score on GREV problems by a larger amount (t(18) = 2.91, p<0.01). GC-long strategy students decrease their GREV score, and indicate no change on GRES problems. This reflects the interaction found for the S-short G-short group compared to the S-long G-long group.

Developing GC-short strategies are good for GREV problems, but bad for GRES problems. Short proofs use situations that are more abstract overall which is better for verbal problems, but not good if the task is to produce a single concrete situation. There is no perceptible effect of different strategies for solving GN-goals. This is perhaps because the different strategies do not relate to use of abstraction or concreteness in representation, but rather refers to whether GN-goals are solved simultaneously (GN-short) or successively (GN-long).

For BW problem types, again GC strategy group seems to reflect differences in score more than GN strategy group (see Figure 7.13), and GC strategy seems to influence changes in BWN score more than GN strategy does. The GC-short strategy group increase their scores on BWS, BWN and BWmodal problems, whereas the GC-long group decrease their scores (t(19) = 2.50, p<0.05; t(19) = 2.22, p<0.05; and t(19) = 3.06, p<0.01, respectively). For BWM problems, all strategy groups seem to increase their scores, meaning that following the HP course is good for all students for these problems.

The GC-short strategy group scored lower initially on BWS, BWN and BWmodal problems (t(18) = 2.76, p<0.02; t(19) = 1.98, p=0.06; and t(17) = 2.99, p<0.01, respectively). Different strategy groups do not score significantly differently on the HFT or the PFT tests.

As with the S-short G-short strategy group, GC-short students tend to score lower on abstract problems before following the HP course, and seem to develop strategies to cope with this weak-
Figure 7.12: Change of scores for GRE problem types by HP GC and GN strategy group.

It was predicted that GCS-long GCM-short students would increase their GRES score but decrease their GREV score, but GCS-short GCM-long students would demonstrate the reverse pattern. A two-way repeated measures ANOVA, with single/verbal GRE problem score nested under time revealed a significant interaction between these two groups (F(1, 10) = 6.16, p<0.05). GCS-long GCM-short students increased their score on GREV problems more than GCS-short GCM-long students (t(10) = 3.55, p<0.01), resulting in them scoring higher on the GREV prob-
Figure 7.13: Change of scores for BW problem types by HP GC and GN strategy group.

Figure 7.14 shows GRE problem means for GCM short/long and GCS short/long strategy use.

Considering strategies in HP as reflecting different use of abstraction/concreteness in representation enables a more detailed characterisation of GCM- and GCS-goal strategies in these terms. GCM-long proofs use representations that are more concrete than GCM-short proofs. GCS-long proofs use representations that are more abstract than GCS-short proofs. Thus, GCM-long GCS-short strategy users are those that have an affinity for concrete representations in their problem solving, whereas GCM-short GCS-long students are ineluctably drawn to use abstract represen-
Figure 7.14: Change of scores for single/multiple/verbal GRE problems by HP GCM and GCS strategy group.

The suggestion that GCS-long GCM-short strategy students develop one strategy good for GREM problems but poor for GRES problems, whereas GCS-short GCM-long strategy students develop a strategy good for GRES problems but poor for GREM problems is only half-supported by the current investigation, as GREM problems are not significantly different for the two groups. However, considering GREV problems against GRES problems reveals that these HP groups seem to develop a strategy that is only appropriate for one problem type, and actually inefficient for solving problems of the other type, where inefficiency is determined by the number of situations used to solve the goal. This supports a taxonomy of problem types that characterises verbal problems
as an extreme version of multiple case problems.

Distinguishing groups by GCS-short/long strategy indicated that the GCS-long strategy group score higher initially on GRES problems than those that use GCS-short strategies (t(19) = 2.12, p<0.05). Also, the GCS-long group end up scoring higher on GREV problems than the GCS-short group (t(21) = 2.04, p=0.05), as a result of improving their GREV score whereas the GCS-short strategy group decrease their score (t(19) = 1.91, p=0.07). The GCM-short group also increase their GREV score whereas GCM-long students decrease their score (t(19) = 2.44, p<0.05). This results in the GCM-short group scoring higher on the GREV post-test score (t(21) = 1.86, p=0.08^3). So strategies that use abstract representations seem to be developed in order to compensate for initial difficulties over abstract GRE problems. Equally, concrete strategies seem to be developed to compensate for low pre-test score on GRES problems.

Students that are adaptive in their strategy for different goal types in HP (the GCS-short GCM-short group) indicate a general deterioration in scores on the GRE from pre- to post-test, though not significantly worse than any other strategy group. This is intriguing, and suggests that a more strait-jacketed approach to problems in HP may be better for transferring reasoning skills to other domains. The GCS/GCM-long group and the GCS/GCM-short group learn something from HP, a general approach that is very useful in some situations, and operable, even if inefficient, in other situations.

For BW problems, all four strategy groups seemed to perform differently (see Figure 7.15). The GCS-long GCM-short strategy group seem to improve their BWS score more than GCS-short GCM-long strategy group (one-tailed t(10) = 1.71, p=0.059), suggesting the reverse of the GRES problem performance. These groups do distinguish change of score on the BWmodal problems, however, with the GCM-long GCS-short group decreasing their scores, but GCM-short GCS-long students increasing scores (t(10) = 2.51, p<0.05).

Distinguishing students by GCM-short/long group revealed several differences. GCM-short students scored lower in the pre-test on BWS, BWN and BWmodal problems than the GCM-long group (t(19) = 2.92, p<0.01; t(20) = 1.75, one-tailed p<0.05; and t(18) = 2.87, p<0.02, respectively). The GCM-short strategy group improved scores on BWS, BWN and BWmodal problems, whereas the GCM-long group decreased their scores on the BWN and BWmodal problems and

^3Significant as a one-tailed test, p<0.05.
did not indicate an increase in score on the BWS problems from pre- to post-test (for change in BWS by GCM-long/short group: t(20) = 2.36, p<0.05; for BWN: t(20) = 1.78, one-tailed p<0.05; and for BWmodal: t(20) = 2.69, p<0.02). GCS-short/long groups did not distinguish any of the BW problem type scores.

Figure 7.15: Change of scores for BW problem types by HP GCM and GCS strategy group.

If the psychometric tests are reflections of flexibility in selecting appropriate levels of abstraction for different problems, then it was predicted that the GCS-short GCM-short strategy group should score highest on the PFT and HFT tests. In fact, the reverse seems to be true. These are students that score low on the PFT and HFT (see Table 7.9), though one-way ANOVAs were not significant (F(3, 17) = 2.00, p=0.15 for HFT score; and F(3, 17) = 0.95, p=0.44 for PFT score). From the table it looks like these measures reflect styles of abstraction use rather than abilities. Students that use abstract representations score highest on both measures, and significantly higher than
the GCS-short GCM-short group (for HFT: t(8) = 2.42, p<0.05; and for PFT: t(8) = 2.06, p=0.074).

<table>
<thead>
<tr>
<th></th>
<th>Mixed GCS-s GCM-s</th>
<th>Mixed GCS-1 GCM-1</th>
<th>Concrete GCS-s GCM-1</th>
<th>Abstract GCS-1 GCM-s</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFT</td>
<td>11.80</td>
<td>13.50</td>
<td>13.57</td>
<td>14.40</td>
</tr>
<tr>
<td>HFT</td>
<td>8.20</td>
<td>13.75</td>
<td>14.29</td>
<td>19.00</td>
</tr>
</tbody>
</table>

Table 7.9: PFT/HFT scores for GCS/GCM strategy groups.

7.5 Discussion

This second HP study succeeded in replicating the aptitude-treatment interactions found in the original HP experiment and the syllogism teaching study. This means that the finer-grained analysis carried out in this study infuses these previous studies – the findings here apply beyond the HP domain.

In contrast to the original study, though, a new range of strategies were identified – the sentential strategy for HP problems had not occurred in the original study. This new strategy sits on a continuum with the strategies previously observed in HP, and the dimension they differ on is the level of abstraction employed in the representations used. The level of abstraction relates to the parallel/abstract and sequential/specific implementations of the ICIA in syllogistic reasoning. As with the syllogistic algorithms, cases are either fully specified in sequential order, or cases have properties that remain unspecified until all are resolved. Using more abstract representations reduces the number of cases to be considered, as the level of “granularity” is greater, but the operations (i.e., possibilities for the “merge” steps in HP) that can be made on these cases is more complex.

The HFT relates to use of these different strategies in HP, reflecting the level of abstraction used by the students. Using more abstraction in the HP representations corresponds with scoring higher on the HFT. The HFT was predicted to be an indicator of preference for different strategies in domains where there is less structure, and more freedom to pursue alternative approaches to problems. HP provides sufficient freedom for the HFT to emerge as a measure of abstraction use.

4Significant as a one-tailed test: p<0.05.
One of the principal aims of this chapter was to “demystify” the GRE test, by interpreting it in terms of the strategy - representation - algorithm framework presented in previous chapters. The extent to which this has been achieved is tied to the success of the taxonomy of problem types as empirically substantive and theoretically useful.

With regard to the GRE problem type classification, there was some support for the distinctions from the clustering analyses – the responses of students reflected the ordering of problems according to the level of abstraction required to solve the problem efficiently. The GRE problem classification also helped to cast light on the ATIs, illuminating who learned what and, to a certain extent, why there were these effects in transfer. Students with previous weaknesses for a particular problem type seemed to develop strategies that dealt with their weakness, but this was an over-compensation which affected their ability to solve other problems effectively. Students that scored low initially on GRES problems, for example, developed a strategy in HP that used more concrete representations (GCS-short GCM-long proofs) and this led to an increase in their GRES problem score on the post-test, but a reduction in their GREV problem scores. The taxonomy of problem types applied to HP and the GRE revealed that there was a general rigidity about the strategies developed by students. Most students used abstract or concrete representations regardless of the requirements of the particular goal. This rigidity was seen in terms of the “holist” and “serialist” strategies used by students in the original HP study – holist proofs relate to use of more abstraction, serialist proofs relate to use of more concrete situations. Pask’s “pathological” problem solvers, then, emerge in the HP course. The problem type taxonomy indicates that the “determinateness” of a problem relates to the effectiveness of alternative strategies, and these strategies can be described in terms of the level of abstraction used in the representations.

However, the success of the taxonomy of problem types is not without constraint. Single case and multiple case problems are shared across the BW and the GRE tests, and there are some differences in response by students to the same type of problem on the different tests.

Multiple case problems are related in these two tests: BWM and GREM pre-test scores correlate highly, but the single-case problems are not so related. Furthermore, students that improve on certain types of GRE problem as a result of following the HP course do not indicate the same change in the BW test. The example of strategy groups on the GCS and GCM HP goal types is a case in point. The GCS-short GCM-long group increase GRES scores but decrease GREM scores. The same group exhibits little change in BWS score and shows an increase in BWM score. The
GCS-long GCM-short group decrease GRES score and increase GREM score, and the same group increases BWS score and increases BWM score.

It was previously mentioned that perhaps constructivism could explain these differences, so to what extent is this true? The hypothesis was that the BW test gives a situation to the student, whereas in the GRE test, the situation has to be constructed from scratch. Therefore, it is plausible that the provision of a situation may encourage strategies that use concrete representations, which could override a student's preference for using more abstract representations. The GCS-short GCM-long group use concrete representations in HP. The BW format is already promoting concrete representations and so little change will be seen in the students that are learning to use such concrete representations for problems that require the construction of situations from scratch. The general increase in BW scores for all students accounts for the BWM increase when GREM scores do not increase, suggesting that solution of BWM problems is not susceptible to strategic variations. The post-test BWM problems may have been easier than the pre-test problems of this type, for instance. The GCS-long GCM-short group use abstract representations in HP and decrease their GRES scores but increase their BWS scores. Again, the promoting of more concrete representations by the BW test means that BWS performance is not impaired by the student learning to use more abstract representations.

The taxonomy of problem types is supported by the empirical results, and helps to indicate the transfer effects from learning a flexible system of problem solving using representations employing various levels of abstraction. How does this relate to the syllogism teaching study? Two psychometric measures overlapped between the HP study and the syllogism study.

In HP, PFT group indicates an effect for single case problems, with PFT-Hi students solving GRES and BWS problems more accurately than PFT-Lo students, whereas for multiple case problems the difference is smaller. PFT group seems to influence effective solution of problems where abstract representations are useful. In the syllogism study, the PFT-Hi group are better at translating out of the graphical, EC representations, which promote a parallel/abstract approach. So, to what extent is the PFT a measure of strategic flexibility? One possibility is that it is a measure of ability to translate between representations which, as previously mentioned in Chapter 6, may be an important aspect of using multiple strategies. The HP study does not indicate that PFT-Hi students are better than PFT-Lo students at other problem types, but nor are they found to be worse. The flexibility hypothesis, then, does not gain support, but it is not disproved either. At the least, the
PFT reflects effective use of parallel/abstract approaches both in HP and in syllogistic reasoning and is silent with regard to sequential/specific approaches.

The GREA subscale was also found to relate to performance on learning methods for solving syllogisms. GREA-Hi students learned better from the EC method, whereas GREA-Lo students learned better from the ND method. The GREA-Hi students are those that are better at both the GRES and GREM problems, and so from the perspective of the taxonomy of problem types, the GREA is a composite measure. The GREA-Hi students therefore present as a group that is flexible and able at different types of problem. The analyses of the "Office" problem (Figure 7.1) and the "Poets" problem (Figure 7.2) earlier in this chapter suggest that different levels of abstraction in representation are required for solving these problems, but that both would be rather difficult and cumbersome if all possible situations are specified uncritically. The original GREA groupings, then, suggest some reflection of the parallel/abstract approach, provided it is borne in mind that such approaches require eventual specification of individuals. How the specification is arrived at is the distinction between the algorithms, and the GREA measures this distinction.

A reanalysis of the teaching syllogism study suggests itself: to assess whether it is GRES or GREM response that predicts response to the two teaching methods. Median splits were performed on the groups according to their scores on the GRES, the GREM and the GREV measures as defined in the second HP study. Two-way ANOVAs with teaching method and GRE problem type group were performed, with errors and interventions on each stage in the method as dependent variable. Table 7.10 shows the interactions. The GREV measure does not discriminate performance on the two teaching methods, this confirms earlier analyses using the "old" GREV subscale measure. Both the GRES and the GREM groups distinguish response to the two methods, though the GRES group distinguishes the translating in stage and the manipulating stages for both errors and interventions. All the interactions are in the same direction as those found for the GREA measure in Chapter 6 – scoring high on the GRES or the GREM problems relates to finding the EC method easier and the ND method more difficult. However, the composite GREA grouping did not distinguish response to the methods for the translating-out stage, but the GRES shows a clear influence here, and there is a suggestion of one from the GREM group. This is the point in the procedure where the selected individual is specified, a process akin to the construction of a single case in the "Office" problem, for example.

Once again, the GREV measure does not predict differences in response to the two methods. This
suggests that one of the other “ignored” dimensions that distinguishes GREV problem types from GRES and GREM problems is perhaps at play. The level of abstraction is somewhat an oversimplification of the potential variation in reasoning approaches. Finding a relationship between the GREV and the ND method in the reverse direction would have strengthened claims for the dichotomy between the parallel/abstract and sequential/specific approaches to problem solving. The ND method does not seem to present the “perfect” strategy for students with a preference for a sequential/specific approach.

In summary, the HP study has shown that use of different levels of abstraction in representations is related to the number of cases constructed in reasoning. These strategies are observable and rigidly applied by students in a number of different reasoning tasks. The final chapter next considers these findings in the context of dual process theories of thinking, devised to explain individual differences in reasoning. The conclusions also extend some of the educational implications of the empirical studies in this thesis.
Chapter 8

Conclusions: theories of reasoning and theories of thinking

Following a summary of the main achievements of this thesis, some of the difficulties and challenges to the account presented are anticipated. Considering alternative accounts for the same phenomena leads to suggestions for future work. These conclusions also contemplate the connection between theories of reasoning and theories of thinking, a connection profitably forged by individual differences research. Dual-process accounts of thinking accounting for different approaches to reasoning are related to the computational styles approach taken in this thesis. Penultimately, the impact of the current studies on educational issues are explored, particularly the implications for the aims and effectiveness of teaching formal problem solving skills. Finally, suggestions for future directions and explorations are made.

8.1 Achievements of the thesis

This thesis has attempted to provide explanations of aptitude-treatment interaction results in terms of the processes related to use of different representations. Such an approach has provided both theoretical and empirical results.

The strategies used for solving problems in a number of domains have been reclassified in terms of the level of abstraction in the representations used in the strategies. There is a general trend
towards developing strategies that use more appropriate levels of abstraction. However, use of these strategies is not entirely due to abilities in developing this appropriate abstraction, but also due to preferences for using different levels of abstraction in representation. This point will be returned to below, when alternative explanations that consider these styles as abilities are considered. The characterisation of the different strategies has been described in terms of the computational processes that are induced by the use of different levels of abstraction in representation, styles that process specific cases sequentially, or more abstract cases in a more holistic manner. Psychometric approaches to strategies and representations have been redescribed in terms of computational styles, which is closer to an explanatory account, and also makes modelling of the different strategies possible.

A further theoretical achievement, with implications for educational research, is the classification of different problem types in terms of the level of abstraction in the representations used to solve them appropriately. This enables a first attempt at characterising what the different problems in the GRE analytic reasoning test are gauging: they probe the flexibility of students to employ different levels of abstraction as appropriate to the problem in hand. This underlies the use of the GREA subscale as a style measure relating to performance in other reasoning domains. Students that are comfortable with use of abstract representations perform better on the GREA test, whereas those that are more au fait with a case-specifying approach do not perform so well on these problem types. Furthermore, such a classification enables the GREV subscale to be considered on the same dimension with the GREA subscale, supporting the presence of both these scales on the same test. Consequently, this means that performing successfully on this test requires using a wide range of abstraction in representations and successfully fitting the level of abstraction to the problem type, which would require metacognitive assessment of the representation to be used before the problem is attempted. Providing an analysis of what the GREA test actually tests is no small achievement. There have been few, if any, studies relating the GREA test to psychometrics (see Enright, Tucker & Katz, 1994), and no process accounts have been attempted, except for a computer program that solves the single case class of problems, developed by Rood (1997). The GREV subscale is similarly underexplored, with descriptive attempts being rare exceptions (Yang & Johnson-Laird, 1999).

A main empirical finding of the studies in this thesis was the generalising of strategies across domains, enabled by considering strategies, representations and algorithms jointly. The same
aptitude-treatment interactions were replicated not only in one domain (HP) with two very different subject populations, but also in a very different reasoning task (syllogisms) with a much shorter teaching intervention. This suggests that the description of the strategies in terms common to both these domains is plausible and, indeed, necessary for a full appreciation of the effects.

Another achievement of the thesis is in terms of the separation of different stages in external representation use. A case was made for the differentiation of translating-in, manipulating, and translating-out of representations when attempting to characterise the use made of external representations to support reasoning. The use of psychometrics supported the separation, but it drives an underexplored field of study into the different influence of the different stages on problem solution, left unexamined by contemporary studies (see, e.g., Scaife & Rogers, 1996; Cheng, Lowe & Scaife, in press). Issues of constructivism (Cox, 1996), for example, relate differently to the various stages in use of representations. What effect does being given representations have on the student’s ability to translate effectively into and out of the representation? What influence on the manipulation of the representations does being given the representation have? These questions are doubtless related but by no means identical, and a breakdown of representation use into (at least) these three stages is a potentially useful addendum to current research into external representation use.

The individual differences in reasoning highlighted by the current analysis indicate that students respond differently to teaching methods that use different representations, and also spontaneously develop strategies that use representations that differ in similar ways to those employed in the teaching. For categorial syllogisms, Ford’s (1995) analysis of spontaneous strategy development gives credence to case-identifiability as the underlying class of algorithms that is being implemented in various ways by students. Furthermore, transfer of skills from one domain (HP) to another (BW and GRE) have been discovered. These skills do not reflect flexible approaches to problems as a result of learning logic, rather they indicate “pathological” tendencies in the student’s strategic repertoire. This inflexibility arises in students that have a weakness for solving certain types of problem, requiring either concrete or abstract representations, and then developing skills to cope with these weaknesses which interfere with problems that they initially had little difficulty with. Such findings have implications for the teaching of logic and strategic flexibility. These will be discussed later in this chapter.
These achievements rest on experiments that are not without difficulties, and the theory behind the interpretations of the results rest on certain assumptions that may not be widely shared. Some difficulties with these accounts are discussed in the next section.

8.2 Criticisms and problems

There are four criticisms that could be levelled at the account of individual differences in reasoning presented in this thesis. The first three are practical concerns, and the last questions the theoretical basis of the account.

One possible criticism is that the experiments reported have all used small populations of subjects. Use of psychometrics, because they tend to be multicomponential, and because test-retest reliabilities are often quite low, will only support theories of task solution if large subject numbers are employed. The lack of correlation between the GRE tests and the HFT, for example, which would be expected to be significant in a large population, point to the weakness of using small groups. The subject group size for the teaching syllogism study is small, as is the group that participated in both the pre-tests and the post-tests in the second HP study. Using subject groups as small as 17 is far below the large numbers used in many psychometric studies.

Relatedly, the second criticism concerns the experimental designs being far from ideal. Different pre-tests are used for the different studies, and no explicit connection between the two reasoning domains is provided. The experimental designs appear rather ad hoc, so comparisons between domains can only be suggestive.

Addressing the first criticism, the use of large subject groups to compensate for the variance in use of psychometric measures was neither necessary nor appropriate in the studies reported in this thesis. This is because large subject groups help to smooth out the unevenness in reasoning performance when related to psychometrics, and this unevenness is the very focus of investigation in this thesis. In addition, the current study aimed to investigate exactly what the psychometrics are testing, by relating them to detailed data available from the studies. Due to the high concentration of data in the studies, it was considered impractical and unnecessary to use large subject populations. Concentrating on strategy variation, rather than crude performance measures means that small subject groups are better fitted. The existence of aptitude-treatment interaction (ATTI) replications in the three studies indicates that the populations are large enough.
to reveal interesting similarities across domains. True generality of the results can be claimed when similar patterns of performance emerge from different subject populations solving tasks in different reasoning domains, and this provides a truer indication of the usefulness of the current approach to elaborating individual differences in reasoning.

The ATI replications are also pertinent to the second practical criticism. These ATIs, which have proved remarkably elusive to attempts at replication (Snow, 1980), show that the overlap between the domains is well-established, and qualitatively similar. However, ideally, a study that directly links the two reasoning domains would provide more complete support for the connection between strategies used for these two domains. Failing that, using all the same pre-tests and post-tests for each of the teaching studies would be next best. The designs of the studies were initially inspired by attempts to relate different representation use to previous studies of strategy change which used only spatial ability measures. The PFT was used in both teaching studies. Unfortunately, the importance of style in strategy change and representation use that was revealed in the syllogism study emerged after the second HP study, so the II task and the serialist/holist questionnaire were not related directly to HP. A study that compared performance on syllogism problem solving to performance on HP would be an interesting investigation, though one perhaps prone to interference between domains. The studies reported are not complete in their overlap, but the areas of commonality are sufficient for the investigations to be more than merely speculative in their contribution to reasoning research.

The third criticism challenges the “subjective” nature of the teaching syllogisms experiment. There was only one instructor that taught the syllogisms, and interventions were made when the subject was judged by the instructor to be having difficulties with the method. There is also a fair degree of subjectivity in the marking of interventions and errors from the videos and transcriptions of the teaching dialogues. It would certainly be useful to have a replication of the teaching experiment using another instructor, or alternatively to construct an automatic tutor for teaching the methods for solving syllogisms. However, justification for the results of the teaching syllogism experiment once again comes from the existence of the ATIs, ATIs that emerge when the instructor is blind to the psychometric scores of the subjects. The replication of similar effects in different reasoning domains suggests that such patterns of interactions are robust, and subjective influences did not affect the results unduly.

The fourth, more theoretical concern, challenges the relationship between the representational
differences observed in the different reasoning domains discussed in Chapters 2 and 4, and the computational account as outlined in this thesis. This relationship is underspecified, underexplored, and furthermore is not the only way of describing the differences. As observations these points are perhaps valid, as criticisms they are not. There are several ways that psychometric accounts can be aligned with computational accounts and this suggests that such alignments are not isomorphic. But this is not to say that such connections are not useful, or that they are theoretically flawed. Good theories provide explanations and predictions for performance as well as descriptions of that performance. Observing the similarity between the ATIs in different domains requires an expression of those commonalities between performance on the various tasks. The sequential/specific parallel/abstract distinction is one such expression.

So is the theory too general to be useful? The results of the empirical studies in this thesis suggest that interpreting reasoning performance in terms of the computational styles is a productive approach. The space of possible strategies is illuminated and a program for effectively intervening in the development and teaching of problem solving skills is generated. A way of assessing the usefulness, generality, and explanatory value of the current approach is to compare it to other accounts that have provided description, classification, and explanation of different strategies in reasoning. Though framed as alternative accounts, without attention to strategies, representations and the underlying processes these accounts are incomplete and may require supplementing with another level of description. In this sense, alternative explanations may be incomplete or complementary accounts of individual differences in reasoning.

8.3 "Alternative" explanations

Another possible criticism of the account provided in this thesis is that it just describes the phenomena at the wrong explanatory level. A better alternative would be to describe different styles of reasoning in terms of the different representations used (e.g., Roberts et al., 1997). However, the starting point of the investigations in this thesis was the inadequacy of describing strategy preference and change in terms of types of representation. Spatial ability measures related to complex problem solving tasks in that subjects that scored high on these measures were better able to develop representations that were more effective for solving the problem. So, the change from spatial to verbal occurs in linear syllogisms, but not for the sentence-picture verification
task, and certainly not for categorial syllogisms.

The next alternative is that the level of abstraction in the representation is the appropriate level of explanation. This is closer to the mark, but it doesn’t capture students’ preferences when different types of representation do not differ in terms of their level of abstraction, and this seems to be the case for categorial syllogisms. Paivio (1986) switched from describing the different representations to consider the processing inherent in using the different representations, and such a reanalysis is necessary for explanations of performance above and beyond descriptions of strategic variation. In HP, the ATI was explained in terms of abstract representations relating to using few cases in proofs and concrete representations relating to strategies that specify many cases. In the syllogism study, the level of abstraction is harder to identify, but the difference is instantiated in terms of how and when the properties of individuals are expressed in the EC and ND methods. In the syllogism domain, the ATI is due to preferences for this general approach, rather than preferences for different levels of abstraction in the representations used.

Rather than challenging the level of description of the strategic differences observed across a range of reasoning tasks, another point of divergence in theory is whether differences are due to styles, or whether such different strategies can reduce to measures of ability. The idea here is that certain strategy patterns emerge due to differing abilities to develop and test alternative strategic approaches to the problem. Roberts and Newton (in preparation) have such a theory: they suggest that strategies develop only when there is a “window of opportunity” for such development. The conditions under which such a window occurs are when the subject has sufficient resources left over from maintaining the representation they are using, and sufficient cognitive load to justify making such changes. So, for example, in the compass directions task, subjects that scored higher on measures of spatial ability were more likely to change from the spatial to the cancellation strategy sooner and these are the students who have better “representational ability”, leaving more resources available for assessment of alternative strategies. However, such a strategy change did not occur for these subjects when pencil and paper was provided, removing the incentive to ease cognitive load during the task.

Such an approach is consistent with accounts that have linked intelligence, or psychometric g, to reasoning ability, where greater intelligence is characterised in terms of greater cognitive resources for problem representation and solution. Kyllonen and Christal (1990) eponymously stated that “reasoning ability is little more than working memory capacity!?” . They found that
general reasoning ability (Carroll, 1989) correlated highly and consistently with general working-memory capacity. Just and Carpenter (1992) put forward a complementary "capacity" theory of working memory that interprets the connection between reasoning ability and the availability of activation as a resource for cognitive processing. Hence, the ability to abstract information is consonant with greater availability of resources, and this is a position put forward by Stanovich (1999). This will be discussed below.

For the studies presented in this thesis, the two "computational styles" (see Table 8.1) discussed could be due to differing abilities for representational ability. One computational style reflects the availability of greater cognitive resources. So, the PFT and the HFT, and the GRE test are measures of "intelligence", or cognitive resources, rather than reflecting cognitive styles. Students that scored higher on any or all these measures demonstrate greater intelligence, and so the strategies that these students utilise are those due to the possession of greater resources. A difficulty for such accounts are that the different "ability" measures do not reflect use of the same strategies. Furthermore, the ATI results present problems for such accounts. Presumably students that have more cognitive resources would learn better from any teaching method, and this was a possibility considered in the syllogism teaching study with regard to the PFT measure. Such a complex pattern of strategy development and the ATIs suggest that the strategies at issue reflect bipolar measures, and hence cognitive styles rather than abilities. Accounts of working memory do have a significant role to play in describing and depicting the processes at play. Absent in the account presented in this thesis is any mention of the engine powering case-storage and inference production. Such an account, however, should be aware that resources are deployed differently according to cognitive style and experiments relating working memory to reasoning ought to take these into consideration.

Studies of the relationship between different components of working memory (see, e.g., Baddeley, 1990) and reasoning performance are a particularly promising starting point. Gilhooly et al. (1993) gave subjects categorial syllogisms to solve in a control condition and a dual task condition. The dual task was one of three secondary tasks: tapping, which occupied the visuo-spatial scratchpad (VSSP); articulatory suppression, occupying the phonological loop; and random generation of numbers between 1 and 10, which interfered with the operation of the central executive component. Of these tasks, only random generation significantly reduced the accuracy of syllogism solving. It was proposed that the articulatory loop has a lesser role in performance,
whereas the VSSP has no involvement at all in the subjects' performance on this task. However, they noted that the presence of dual tasks forced subjects to change strategy for solving the problems. It seemed that most subjects were employing an atmosphere or a matching strategy while solving the problems, with only a few working on the "logic" of the problem. In the dual task conditions, subjects switched to less sophisticated strategies. Toms, Morris and Ward (1993) also found that, for conditional reasoning, occupying the VSSP and the articulatory loop did not interfere with reasoning performance, but central executive dual tasks did. Vandierendonck and De Vooght (1997) found that all three components of working memory were involved in linear syllogistic reasoning, whereas Oberauer et al. (1996) found that the VSSP did have a role in syllogistic reasoning. These somewhat conflicting results provide further evidence that the different strategies that subjects are employing on the tasks must be taken into consideration. Relating dual task conditions to the use of taught strategies such as the EC or ND methods for solving syllogisms, or classifying students on a pre-test into those that use spatial or verbal strategies on a given task (Ford, 1995) may assist in assessing the variation and change in the resource deployment of working memory components in reasoning tasks (Shah & Miyake, 1996).

A further possibility for redescribing the results presented in this thesis comes from the heuristics and biases literature. Recent work in this field has offered "dual process" accounts of reasoning, which postulate two different systems, or processes, that individuals can apply to problem solving situations. These theories are discussed in the next section, relating theories of reasoning to theories of thinking. These dual process accounts provide a useful generalisation of the current investigation, and also link the current theory of individual differences in strategy and representation use to heuristics and biases in reasoning.

The dual process accounts are relevant to the account of use of different strategies, representations and algorithms in reasoning because they indicate a similar dichotomy of approaches. The extent to which the individual is prone to contextualisation of information – a central theme of dual process accounts – is a symptom of the vulnerability of the representational system to such influences. The point, in summary, is that contextualisation is more likely to occur with sequential/specific approaches as contextualisation enables individuals to be fully specified (even if these are at the expense of covering the space of possibilities). This point will be elaborated further following a review of dual process accounts of reasoning.
8.4 Theories of reasoning and thinking

This thesis has presented arguments and experiments in support of an account of reasoning that details two learning styles that are based on different computational styles of processing. These two styles are manifested with a number of related properties, and have been shown to relate to different psychometrics as shown in Table 8.1.

This is reminiscent of other dual-process accounts in reasoning. Three theories in particular are closely related to the account presented in this thesis. They differ from the current account in that each makes moral claims about the reprehensibility of using one or other process for reasoning tasks. The two reasoning styles depicted in the current study are essentially different ways of approaching tasks, which may be more or less appropriate according to task requirements.

In order to account for the apparent irrationality in reasoning tasks of highly intelligent individuals, Evans and Over (1996) formulated a distinction between two forms of rationality. Rationality₁ is attuned to actions that are efficient and general for achieving personal goals. Rationality₂ has a more objective bent, producing actions that are based on a reason sanctioned by a normative theory. Rationality₁ is achieved via tacit cognitive processes whereas rationality₂ is served by explicit processes. They note that "our thinking is highly focused on what are subjectively relevant features in the task information and from memory" (p.143), and this form of processing is very quick, powerful, preconscious, parallel and implicit. Furthermore, this form of processing is exhibited as "biases" in reasoning in the Linda problem (see below), or in the four-card task, for example. However, it is coupled with an explicit reasoning system, and the operation of the two systems combine in determining actions. The operation of these two sys-
tems is interactive, rather than sequential or conflicting. Inferences and responses can result from the operation of the implicit system with the subject giving an answer that "feels right", or the subject may make a conscious effort at explicit reasoning and so exhibit System 2 processing. When the explicit system is employed, it is still prone to the limitations imposed by the implicit system:

For example, on the selection task not only is people's conscious thinking restricted in general to the matching cards, but their thought about the consequences of turning the cards is also limited to the matching values on the other side (Evans & Over, 1996, p.146).

Sloman (1996) also distinguishes two systems of reasoning. The first, related to Evans' "heuristic" system, is based on associative processing: "associative reasoning inherits a property of associative systems: It computes on the basis of similarity and temporal structure" (Sloman, 1996, p.4). It has the properties of reflecting "similarity and contiguity", operating "reflexively", is parallel, and is generally unconscious processing. In contrast, the rule-based system is defined in terms of its having the properties of "productivity", "systematicity", and operating sequentially. Evidence for this distinction is found when subjects hold simultaneously contradictory beliefs, each generated by the separate systems. Tversky and Kahneman's (1983) "Linda problem" is a prime example of a task where this occurs. Subjects were given the following paragraph:

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Then, subjects are asked to rank eight statements about Linda according to their probability. These included the following sentences:

Linda is a bank teller.
Linda is a bank teller and is active in the feminist movement.

The latter statement is a conjunction of the former with a further condition, so the probability of this statement being true has to be less than or equal to that of the former statement. Most subjects judge the latter statement to be more likely. Sloman interprets this as the operation of the associative reasoning system. Applying probabilistic theory, as done by the experimenters, for example, gives a conflicting response, exemplifying the operation of the rule-based system.
Gould's introspections on this problem reveal the contradictory beliefs generated by different systems: "I know that the [conjunction] is least probable, yet a little homunculus in my head continues to jump up and down, shouting at me—'but she can't just be a bank teller: read the description'" (1991, p.469, quoted in Sloman, 1996).

There are two criticisms levelled at Sloman's dichotomy in a commentary by Gigerenzer and Regier (1996), and these are pertinent for all such attempts to divide nature (or nurture) into two. The first criticism is that Sloman's distinction is inchoate and blurred. The second is that as more and more distinctions are superimposed, the less testable becomes the division as clusters of properties rather than single properties become the issue in question. Two methods of addressing these related problems are to, as Gigerenzer and Regier suggest, turn divisions into operable computational models. Alternatively, properties of the data that reflect use of different reasoning systems must be isolated. This latter course has been the fruitful approach of Stanovich (1999) and Stanovich and West (in press).

Stanovich classifies the dual-process theories of Evans and Over (1996) and Sloman (1996) as being cases of the same distinction. He terms the two reasoning processes System 1 and System 2. System 1 is fundamental, automatic, heuristic (Evans, 1989), implicit, associative (Sloman, 1996), and holistic. System 2 is controlled, analytic, explicit, and rule-based. Further, System 1 is inclined to impose a context when, or perhaps in order to, interpret problems. Applying conversational conventions to problems where this is inappropriate is a feature of using System 1. System 2, in contrast, is characterised by decontextualisation of information. Decontextualisation has frequently been conflated with the ability to abstract. Donaldson (1978) located the ability to disembed from the immediate surroundings as the essence of learning to use abstract rule systems, and consequently the development of higher intellectual skills. Denny (1991) defines decontextualisation as "the handling of information in a way that either disconnects other information or backgrounds it" (p.66), and this resonates with the field-independence/dependence cognitive style. This link between reasoning systems and decontextualisation means that testable predictions can be made about the extent to which an individual reasons with each of the two systems, with respect to the level of decontextualisation that is employed.

This propensity for the context to be pervasive in information processing is termed by Stanovich the "fundamental computational bias" in reasoning. The main processing tendencies covered by this bias are the tendency to adhere to Gricean conversation principles, to use prior knowl-
edge, and to reason enthymematically (1999, pp.192-193). This explanation serves to explain the conjunction fallacy in responses to the Linda problem, even among the experimenters that have pondered the problem there is still a bias towards responding with “the social and pragmatic contextualization necessary to support a Gricean mechanism of intention attribution” (Stanovich, 1999, p.197).

What attempts have been made to explore individual differences in dual-process systems of reasoning? In a large-scale series of studies, Stanovich and colleagues have related performance on a number of reasoning tasks with measures of “thinking dispositions” and “cognitive ability” (Stanovich & West, in press). They found that normative responses to categorial syllogisms and the four-card task were correlated with SAT scores (p<0.01). They took this to indicate that, to a certain extent, computational limitations determine response. However, there was no neat connection between responding “logically” and scoring high on the SAT. Stanovich and West (1998) note that several students that score low on the SAT respond correctly to all syllogisms, and to the four-card task, whereas some students that score high on the SAT do not make the logical responses to these tests. When SAT scores are partialled out, there remains a correlation between performance for the four-card task and for syllogisms (p<0.01). This suggests that a decontextualised response to these tasks is a matter of style. Stanovich and West (1998) conducted a further study to measure responses to several items that purported to assess thinking dispositions with responses to reasoning tasks. Such items questioned the students' openness to operate in a decontextualised manner, so, for example, they were questioned on the extent to which they may remain undecided about an issue, about whether there are moral absolutes, or whether they can distinguish beliefs from debate on issues. As Stanovich (1999, p.191) notes, these items “are indirectly tapping the processing style of cognitive decontextualization”. Strong correlations were found between a composite score of these thinking dispositions and syllogistic problem solving (p<0.01), as well as with a host of other reasoning tasks.

For a study of belief bias effects in syllogism solution, Sá, West and Stanovich (1999) gave their subjects problems with conclusions the content of which were either true if the syllogism was invalid, or false if the syllogism was valid. They found that a high cognitive ability score related to responses that ignored the real-world content of syllogisms, and focused on the form of the argument.

Stanovich concludes that System 1 is more basic, and individual differences in response to rea-
soning tasks are due to differences in ability to decontextualise information and apply System 2 processes to the given task. Ability to do this is related to psychometric g, but is not wholly determined by cognitive ability, and can be equally well determined by questionnaires about individuals’ dispositions to decontextualise information.

The dispositional approach to applying the different reasoning systems is vindicated by the results of variation in strategy use in this thesis. The influence of g as an indicator of the use of the different computational styles, however, is implausible. Reasons for rejecting theories that “intelligence” predicts strategy use were presented earlier in this chapter. To reiterate, “ability” measures such as the GRE test, seem to rather reflect styles of processing. The prevalence of ATIs testify to this and present a challenge to accounts that posit one strategy as more fundamental or less sophisticated than another.

The question as to whether the two computational styles reflect different applications of System 2 processes can be addressed by examining the hallmark feature of the System 1 and System 2 processes identified by Stanovich: the extent to which information is decontextualised in problem solving. There are explicit tests used to predict decontextualisation that were used in the studies reported in this thesis. The HFT has been used to indicate the extent to which information about diagrammatic figures can be accessed independently from its background or context. So, scoring high on the HFT relates to a propensity to decontextualise information. The immediate inference task can also be used to gauge the extent to which information from quantifier interpretation is decoupled from conversational contexts (Newstead, 1995). So, certain patterns of response on the immediate inference task will reflect the extent to which the individual decontextualises information.

There was no effect of HFT group found in the teaching syllogisms experiment. However, this measure did prove fruitful for the HP study. Scoring high on the HFT relates to using more abstract representations in HP, in particular on Question 4, HFT-Hi students represented problem information in the sentential modality when problems were presented with a graphical situation (see Chapter 3). Expressing graphically presented material sententially is an indication of decontextualisation of information. Sá et al. (1999) contend that applying System 2 reasoning is the ability to abstract, rather than the ability to solve abstract problems, and that this is the focus of individual differences. This seems to be precisely the behaviour of the HFT-Hi group for HP problems.
The HP studies indicate that students learned differently from HP according to their "disposition" to decontextualise information. It is not the case that students learn to decontextualise information as a result of following the course, however. The strategies used by different students varied in terms of the level of abstraction used, and in some cases the students ceased to use abstract representations, and this affected their performance on the determinate GRE problems. There is therefore no hint that students learn to operate System 2 decontextualisation as a result of following the HP course, which might be thought of as an aim of formal logic courses. This is no negative result, however, rather the landscape of learning is found to be far more varied than might have been anticipated. The styles are the issue, and these predict what is learned from the teaching interventions.

The GRE as a transfer task is particularly interesting in that it requires a selection of both cooperative or "formal", or normative, communication modes. In the "Office" problem (Figure 7.1), for example, when told that there are six office workers, it is important to assume there are six and only six office workers otherwise it is going to be very difficult to get going at all. But when told in the "Poet's" problem (Figure 7.2) that some of the students who like Eliot's poetry also like Auden's poetry, it is important to assume that possibly all the students that like Eliot like Auden also. Hence, scoring well on the GRE reflects selective decontextualisation. Scoring less well on the GRE test, then, may be a reflection of an inability to select which implications to maintain. Having a disposition to decontextualise all information may then be an impediment to solving certain GRE problems.

A hallmark of using reasoning System 1 is the use of background knowledge, presupposition and conversational implicature, such as Grice's principles of communication (1975). Contextualisation encourages specification of properties, the focus on individuals perhaps at the expense of considering all alternatives. The II task can be used to reflect the extent to which information is contextualised, when no apparent context is present. Applying a context for quantifier interpretation is likely to encourage rash responses. If "Some As are Bs" is contextualised as "Some animals are dogs", then this will lead to the rejection of "All As are Bs", when "can't tell" is the normative response. Students that respond rashly are those that accept no-conversion, perhaps as a result of contextualising the statements. Those that do not accept no-conversion tend to be hesitant in their response, which suggests that no contextualisation, or specification of individuals occurs for these students. Convertors prefer the parallel/abstract approach, Non-convertors
prefer the sequential/specific approach. Though System 1 processes in Stanovich's account seem to have the holistic and parallel properties of a Style 1 approach, this is misleading, as the primary distinguishing feature of the two styles is in terms of the extent to which properties of individuals are specified or left abstract. System 1 therefore relates more closely to Style 2 and System 2 to Style 1. Furthermore, Convertors seemed to score lower on the HFT than Non-convertors, and performed better on the GREA subscale, suggesting that they were skilled in decontextualising information (see Chapter 6).

Newstead (1995) explicitly identified Gricean errors in the II task as rejecting “All As are Bs” given “Some As are Bs”, and accepting “Some As are not Bs” given the same statement. Subjects in the syllogism teaching experience responding in this way did not have lower scores on the HFT ($t(15) = 0.98$, ns) though it was in the expected direction. Subjects that made Gricean responses, however, did score lower on the GREV subscale (3.29 compared to 5.29, $t(15) = 2.47$, $p<0.02$) and approached significance with the GREA subscale (7.00 compared to 8.64, $t(15) = 1.57$, one-tailed $p=0.07$).

A close link between the Sá et al. (1999) study and theories of separate processing in the two hemispheres emerges in a study by Deglin and Kinsbourne (1996). In a perhaps unique experimental design, they produced syllogisms with premisses that were not true and gave them to patients undergoing unilateral ECT treatment for schizophrenia or manic-depressive psychosis. Unilateral ECTs suppress activation in one hemisphere, and the patients in this study underwent both right-hemisphere and left-hemisphere suppression as part of their treatment, and syllogisms were given to patients during suppression of each hemisphere. After left-hemisphere suppression, rejection of false premisses occurred much more often than in a control condition, and subjects refused to use the premisses in order to reason. Under right-hemisphere suppression the number of answers according to the form of the premisses increased:

...subjects' attitude to false premises changed radically. The subject who showed pronounced emotional responses to false premises under left hemisphere suppression, now performed formal-logical operations quite calmly, with confidence, and remained unmoved by the absurdity of the information offered by the premisses (p.300).

Deglin and Kinsbourne take the results to indicate that the left hemisphere is involved in formal, logical processing, whereas the right hemisphere operates on "empirical" information. They consider this a demonstration of the hemispheric basis to Goldstein's abstract/concrete attitude
distinction, and equally a demonstration of the complementary thinking styles of context-bound and context-free mechanisms for reasoning. This study can equally be given a "decontextualisation" gloss, with the right hemisphere operating contextually, and the left hemisphere attempting to process without the influence of context. The ability to decontextualise information calls for the dominance of the left hemisphere processing over the right hemisphere. Individual differences in reasoning can thus be seen to be relate to hemispheric operation and interaction, and this brings the discussion almost a full circle from the original formulations of the analytic/holistic distinction.

Decontextualisation is a related issue to the studies presented in this thesis, but the computational account is not reducable to the dual-process framework. The ATIs in this study indicate that the two computational styles are not grounded in reasoning systems that differ in terms of how "fundamental" or "primitive" they are. The g component is not a useful way of describing these styles. Rather, the computational styles indicate alternative means to making computation tractable in a variety of situations. Some convergence with the patterns of inference called upon by contextualised as opposed to decontextualised problem solving would be a useful approach to describing the thinking dispositions of Stanovich and West’s (in press) account.

8.5 Educational implications

One of the aims of this thesis was to explore individual differences in the way students respond to learning from formal systems. So what educational implications can be garnered from the empirical studies?

The ATIs show that, when teaching formal systems, the student’s profile is crucially important in terms of what will be learned from different representations. Three questions arise with respect to applying this research in educational programs:

- Should educational programs be designed to address students’ strengths or their weaknesses? So, should students that find abstract representations difficult to use be presented with concrete representations? Relatedly, should students that have a preference for case-limiting strategies be trained to use inference-limiting strategies? Should learning be (maximally) difficult for students?
• Ought the focus to be training in cognitive strategies or metacognitive control? Will more effective interventions be as a result of teaching strategy assessment rather than strategy width? How can flexibility be taught?

• Relatedly, how can transfer of problem solving skills be better facilitated?

An answer to these issues can, to a certain extent, be approached by the results of the current studies, but testing the efficacy of certain interventions requires future work in order for necessary empirical justification to be provided.

Some students seemed to learn more quickly from presentations that benefitted parallel/abstract approaches, whereas other students learned more quickly from representations geared towards a sequential/specific approach. If a student had their style matched to a method, then the first HP study indicated that they transferred skills better to transfer tasks, if there was a mismatch then they became worse at problems in transfer tasks. Putting the syllogism and the first HP experiments together indicates that teaching matched methods may be better for students: they acquire them more easily and they seem to transfer skills to other tasks better. It is predicted that students with a preference for parallel/abstract methods would better transfer the skills they learn from the EC method, whereas sequential/specific students would transfer skills from the ND method. This is a matter for further investigation.

One explanation for an aptitude-treatment transfer interaction may be that matching methods to styles means more cognitive resources are freed up when learning the method. This means that alternative strategies can be pondered (Roberts et al., 1997) and extra resources are available for metacognitive control (Crowley, Shrager & Siegler, 1997). Another explanation may be that when there is a match between style and method this means that the procedure can be learned at a deeper level. Some of the subjects with a sequential/specific preference in the syllogism teaching study seemed to take a procedural approach to the EC method, for example, and some of the parallel/abstract preference subjects had no sense of the “workings” of the ND method. Without a sense of the way the method works there can be no generalisation of the skills acquired in learning, a lesson suggested in the analogical literature (Gick & Holyoak, 1983).

Future empirical work that would cast light on these questions would involve teaching students methods that either matched or mismatched their cognitive style. This could be done with the EC/ND methods for solving syllogisms, or with HP courses that restricted the types of abstrac-
tions that could be made: so the student would be taught to use abstract representations, or
congrrect representations. Then the transfer of skills to other problem solving domains could be
assessed.

The results of the second HP study are pertinent to the second question raised above. The study
indicated that certain students learn strategies that compensate for their initial weaknesses: they
get “locked in” to certain ways of responding. This has the result of making performance on other
types of problem worse: they learn a strategy which is good for addressing their initial weakness,
but interferes with performance where this strategy is not appropriate. Such strategies are either
pathologically abstract or concrete, and these strategy types can be interpreted in terms of pref-
erences for different processes. HP presents sufficient flexibility such that both computational
styles can be served by the domain. These students had a good idea about their weaknesses, but
seemed to learn only one approach to problems. The flexibility of the domain was not always
reflected by flexible use of different levels of abstraction. Some students, however, did seem to
learn a flexible approach to problem solving. This is particularly clear in the first HP study, where
those that scored highest on the GREA subscale were those that used strategies appropriate to
different tasks. This is also confirmed in the second HP study, with the GCS-short GCM-short
group.

The conclusions, then, are that some students have a propensity for learning a single strategy
for problem solving, regardless of the appropriateness of this strategy. This is a reflection of be-

haviour in the reasoning domains discussed in Chapter 2 where some students do not progress
with strategy development. The HP studies show that students are good at addressing their
weaknesses, but they do this at the expense of their initial strengths, under certain conditions.
The problem becomes coaxing the pathological students to be more flexible in their strategy de-

ployment, indicating the inappropriateness of certain approaches to problem solution. How this
is to be achieved is a matter of empirical investigation initially. Flexibility is more likely to occur
if methods are matched to students’ computational styles, because then a procedural approach to
problems is less likely to occur, as are attempts to apply inappropriate representations learned in
different contexts. The current studies do at least give a foot-up to attempts to facilitate flexible
strategy use by identifying student styles and teaching domains that are likely to be susceptible
to particular patterns of strategy inflexibility. Thus, the students that initially score low on single-
case GRE problems are more likely to develop strategies that utilise a lower degree of abstrac-
tion in representation. Students that score higher on these problems are more likely to develop strategies using more abstraction in representations. Training programmes can thus be tailored according to the styles of the student, attempting compensating exercises for the propensity to deploy only one type of strategy.

Another alternative would be to teach students different strategies explicitly, and observe the repertoire of their use. This may again result in inflexible strategy use, or utter demoralisation, or more optimistically it may contribute to deeper understanding of the task. If this turned out to be the result, then multiple representations would then promote flexible approaches to reasoning.

The issue of teaching transfer of skills across domains has been a cause of concern. Stanovich’s (1999) fundamental computational bias indicates that what usually happens is that performance and learning occurs in a very context-specific area, and such skills do not transfer from task to task. Perkins and Salomon (1988) consider attempts for transferring skills between tasks. They argue for the importance of including “bridging” between tasks during training. One contribution this thesis has made to the issue is to indicate occurrences of strategy or skill transfer between domains when these are not explicitly taught. This is true both for near-transfer tasks from HP, such as the BW, which might be anticipated by Perkins and Salomon, but also in tasks where the similarity between domains is less clear, such as the GRE test. The studies in this thesis also explore the cognitive profile of students that present with effective transfer. Understanding when and why such transferences occur provides a better means for structuring programmes designed to facilitate transfer of problem solving skills. A study by Zhang, Johnson and Wang (1998) examined the transfer of skills across different presentations of the same problem, conjoining issues of representation with skill transfer.

Zhang, Johnson and Wang (1998) examined the transfer effects of different representational formats for the tic-tac-toe algorithm. They focused on three of the representational schemas used in Zhang’s (1997) study: lines, numbers, and colours. The line presentation involves a spatial array of 9 positions, where the winning pattern is to select three positions in a line. The number version requires the selection of numbers from 1 to 9, the winning pattern being to get 3 numbers that add up to 15. The colours method represents the different items in terms of clusters of 2, 3 or 4 differently coloured circles. The winning pattern here is to select three groups which all contain the same colours. Students were required to play a computer, which meant that their task was to gain a draw. Subjects were exposed to one version of the task, until they gained draws ten times
consecutively. Then they were given one of the other representations of the problem.

Zhang et al. found that transfer from the number presentation to the colour presentation was positive, whereas transfer from lines to numbers was negative. Thus, students were quicker at achieving success at the colour representation following the number representation, and slower when given the colours following the line presentation. However, the strategies used by students on the different presentations varied, and this seemed to reflect different levels of comprehension of the abstract algorithm. The task can be solved just by choosing any two even numbers every time. Some subjects solved the task by choosing the same two numbers every time – this is termed the “fixed-number” strategy, and does not demonstrate explicit understanding of the two-evens algorithm. Zhang et al. found that students utilised the fixed-number strategy to solve the number presentation much more frequently. For the line method, students discovered that two evens had to be selected, but were more flexible in their actual choices, so they learned a more general strategy. These differences in strategy use affected ease of transfer across representations, as after learning from the number presentation students were looking for a specific fixed number response which could be found more quickly than the more general two-evens strategy in the colour presentation. Students who had seen the line presentation initially had learned the more general strategy which took longer to find in the colours. It was the line students that learned the more general strategy, according to the authors. Zhang writes “task performers usually only deal with the specific representational and implementational contents in which the abstract structures are only implicitly embedded” (1997, p.184).

This is an echo of the discussion on requiring a match between representations and the styles of the student in order to promote effective transfer. There is no individual differences account in Zhang et al.'s work, but ATIs corresponding to those found in the syllogism teaching and the HP studies would be anticipated. The different levels of abstraction in the three forms of tic-tac-toe representations are likely to relate to the different computational styles of students. Learning generalities across domains follows from learning the structures that are implicit in a particular domain, and this is most likely to be achieved in cases where aptitudes and treatments are matched.
8.6 Future work

Proposals for a number of extensions to the current studies have been suggested throughout this chapter. One of which involved generalising the instruction in the syllogism teaching, ideally by constructing an automatic tutor. This could then be used to compare the EC and the ND methods with the adapted-EC method, discussed in Chapter 4. Proposals were also made for studies to improve flexibility of strategy use and effective transfer of skills across domains by adapting teaching methods according to the computational style of the student. Some further work to tidy up the studies by applying the same tests in each experimental domain would also be useful.

The groundswell of research in individual differences in reasoning has primarily been concerned with the dual-process accounts of reasoning. An alignment of the "thinking styles" measures from this literature with the computational approach taken in analysing the current studies would be beneficial for forging generalisable and converging accounts of reasoning performance.

A finer-grained analysis of the different ways the working memory components may contribute to the different strategies would be a useful next step on the way to modelling the computational styles. Particularly interesting would be the different deployment of resources at different points in the task – are there different loads related to parallel/abstract cases as opposed to sequential/concrete cases, for instance. Levesque's (1988) suggestions predict that representing and operating on abstract cases is a different type of effort to that required by sequentially processing fully-specified cases. Working memory models would be complementary to rational analysis accounts of the two computational styles.

There is a fine-line between distinguishing the pedagogical issues of how and what to teach, from cognitive theories of how and what is learned, from philosophical issues of what logic is and ought to be. The aims of teaching logic seem inextricably tied to the aims for studying the nature and styles of reasoning. This link is furthered by the importance of individual differences in relation to both. For teaching logic, the studies in this thesis indicate categorically that what representations are used for reasoning are crucial for determining what skills are transmitted to the student, and this is a function of their aptitude or propensity for certain styles of processing. For studying the nature of reasoning, individual differences prove to be a crucible (Underwood, 1975) for investigating the space for variation in the operation of the dual processes. The intended interpretation of the dual processing in this thesis is one that combines accounts in terms of
representation, strategy, and computational style, but equally the reader can subscribe to any one of a number of dual-process accounts. Each seems to describe the same dichotomy, and each is converging on the evidence for combining the output from each processing system. Deciding between them depends on selecting the best in terms of generality and power for prediction. A computational account, no matter how "abstract", better serves this aim.

The purpose of logical exposition, for Kant, is to postulate an organon, or a processor, as well as a description of the logical steps themselves, and thus logic necessitates a description of the laws of the psyche – a psychology:

Now if logic is a mere theory of the conditions under which a cognition is perfect according to the laws of the understanding and of reason, then it is not a theory of execution; it would be a theory but not an organon (The Blomberg Logic, part 26).

Aligning psychometric and computational accounts of reasoning provides a means for approaching this ideal.
Glossary of abbreviated terms used, with page of first mention.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATI</td>
<td>Aptitude-Treatment Interaction (p.179).</td>
</tr>
<tr>
<td>BW</td>
<td>Blocks World test (p.59).</td>
</tr>
<tr>
<td>BWM</td>
<td>Multiple-case problem on the BW test (p.145).</td>
</tr>
<tr>
<td>BWmodal</td>
<td>&quot;Might follow&quot; problem on the BW test (p.146).</td>
</tr>
<tr>
<td>BWN</td>
<td>Non-consequence problem on the BW test (p.145).</td>
</tr>
<tr>
<td>BWS</td>
<td>Single-case problem on the BW test (p.145).</td>
</tr>
<tr>
<td>EC</td>
<td>Euler's Circles method for solving syllogisms (p.77).</td>
</tr>
<tr>
<td>EFT</td>
<td>Embedded Figures Test, used to test FID (p.48).</td>
</tr>
<tr>
<td>FID</td>
<td>Field-independence/dependence cognitive style dimension (p.47).</td>
</tr>
<tr>
<td>G-goal</td>
<td>Graphical goal in a HP problem (p.148).</td>
</tr>
<tr>
<td>GC-goal</td>
<td>Graphical Consequence goal in a HP problem (p.149).</td>
</tr>
<tr>
<td>GN-goal</td>
<td>Graphical Non-consequence goal in a HP problem (p.148).</td>
</tr>
<tr>
<td>GCS-goal</td>
<td>Graphical Consequence single-case goal in a HP problem (p.149).</td>
</tr>
<tr>
<td>GCM-goal</td>
<td>Graphical Consequence Multiple-case goal in a HP problem (p.149).</td>
</tr>
<tr>
<td>GRE</td>
<td>Graduate Recruitment Exercise, test of analytic reasoning ability (p.53).</td>
</tr>
<tr>
<td>GREA</td>
<td>Analytic reasoning subscale of the GRE (p.53).</td>
</tr>
<tr>
<td>GREM</td>
<td>Multiple-case problem on the GRE test (p.142).</td>
</tr>
<tr>
<td>GRES</td>
<td>Single-case problem on the GRE test (p.140).</td>
</tr>
<tr>
<td>GREV</td>
<td>Verbal reasoning subscale of the GRE (p.53).</td>
</tr>
<tr>
<td>HFT</td>
<td>Hidden Figures Test, used to test FID (p.48).</td>
</tr>
<tr>
<td>HP</td>
<td>Hyperproof (p.57).</td>
</tr>
<tr>
<td>ICIA</td>
<td>Identify Critical Individuals Algorithm (p.78).</td>
</tr>
<tr>
<td>II</td>
<td>Immediate Inference task (p.89).</td>
</tr>
<tr>
<td>ND</td>
<td>Natural Deduction method for solving syllogisms (p.78).</td>
</tr>
<tr>
<td>PFT</td>
<td>Paper Folding Test, used to test spatial ability (p.30).</td>
</tr>
<tr>
<td>S-goal</td>
<td>Sentential goal in a HP problem (p.148).</td>
</tr>
<tr>
<td>SPQ</td>
<td>Study Preference Questionnaire, used to test serialist/holist preference (p.112).</td>
</tr>
</tbody>
</table>
References


Bruner, J.S. (1957). Going beyond the information given. In J.S. Bruner et al. (Eds.). *Contemporary Approaches to Cognition*.


Lawrence Erlbaum Associates.


