Exploring the Determinants of Dual Goal Facilitation in Wason's 2-4-6 Task

Maggie Gale

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Derby
School of Education, Health and Sciences

August, 2008
Acknowledgements

I wish to thank my supervisor Dr Linden Ball for the encouragement, guidance and expertise he has given freely throughout my studies. This thesis has been in progress for a very long time; he has been extremely patient, understanding and forbearing throughout. I would also like to thank Dr Steve Croker for his helpful insights on drafts. I am also appreciative of the help I have received from other colleagues both at the University of Derby and in the wide academic community who have helped by discussing and exploring ideas with me.

I am indebted to all who participated in my experiments; there have been very many people who have willingly given up their time over the years to take part. Their help is much appreciated.

I would also like to express thanks to my husband, Nick for his help and support both financial and emotional, without whom I would not have been in a position to study for my doctorate. Finally I wish to thank my daughters for their enthusiasm, and their blind faith that I could get to the end which they supplied with humbling certainty.
Abstract

The standard paradigm for exploring hypothesis testing behaviour is Wason's (1960) rule discovery task, which exists in two variants: the standard single goal (SG) task, and the logically identical dual goal (DG) form. Despite the close similarity of the two forms of the task, the reported success rates in the two variants vary considerably, with approximately 20% of participants successfully solving the SG variant compared to over 60% correctly announcing the rule in the DG form. It was this disparity between the patterns of performance across the two versions of the task which formed the impetus for this thesis, as it was felt that an explanation for the facilitatory effect of DG instructions would lead to insights into the poor performance in the SG form. Several competing contemporary accounts of the effect are introduced, and predictions derived from them empirically tested across a series of seven experiments. Data analyses showed that no single contemporary theory could provide a wholly adequate account of the DG facilitation effect. However, these analyses led to a novel observation: that it is the production of a contrast class triple which appears to be the key predictor of success on the task, and furthermore, that the DG variant of the task promotes the generation of such a triple. Support for the “contrast class” account of the DG effect was provided by direct manipulation of the information provided to participants. A theoretical account of the critical role of contrast class cue information is developed in the thesis by situating the account within a proposed extension to Oaksford and Chater's (1994) “Iterative Counterfactual Model” of hypothesis testing. It is further suggested that rather than providing mutually exclusive accounts of the DG effect, competing theories (e.g., Vallée-Tourangeau et al.'s, 1995, triple heterogeneity theory, and Wharton et al.'s, 1993, information quantity theory) could be subsumed within this new model, which would then reflect a process whereby participants' strategies change and develop over the course of the hypothesis testing session. Finally, it is suggested that findings from this thesis can be accommodated more generally within Evans' (2006) “hypothetical thinking framework”, and thereby within contemporary dual process accounts of reasoning.
Contents

Acknowledgements ..................................................................................................... ii
Abstract .................................................................................................................. iii

Contents .................................................................................................................... iv
List of Tables ............................................................................................................ ix
List of Figures .......................................................................................................... xii

Chapter 1 Introduction to Hypothesis Testing Behaviour ........................................ 13
  1.1 Introduction ............................................................................................. 13
  1.2 Review of the Literature .......................................................................... 14
    1.2.1 Scientific Reasoning and the “Problem of Induction” ...................... 14
    1.2.2 The Nature of Hypothesis Testing Behaviour ................................ 18
    1.2.3 Wason’s (1960) 2-4-6 Rule Discovery Task .................................. 24
    1.2.4 The Dual Goal Paradigm ................................................................ 30
    1.2.5 Theories of DG Facilitation in the 2-4-6 Paradigm ...................... 32
  1.3 Summary ................................................................................................. 36

Chapter 2 General Methodology and Analysis ................................................... 39
  2.1 Introduction ............................................................................................. 39
  2.2 Scoring Success on the DG Task ............................................................ 39
  2.3 Coding of Triples for Analyses of Hypothesis Testing Strategies .......... 39
  2.4 Restricting Opportunities for Rule Announcement ................................. 40
  2.5 Reporting of Effect Sizes ...................................................................... 41
  2.6 Path Analysis and Logistic Regression ................................................... 42

Chapter 3 Experiment 1: Comparing Positivity Bias and Goal Complementarity Theories ........................................ 43
  3.1 Introduction ............................................................................................. 43
  3.2 Method ................................................................................................. 44
    3.2.1 Participants ..................................................................................... 44
    3.2.2 Design ............................................................................................ 45
    3.2.3 Procedure ..................................................................................... 45
  3.3 Results ..................................................................................................... 46
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.1</td>
<td>Solution success across conditions</td>
<td>46</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Quantity and variety of triples generated across conditions</td>
<td>46</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Presence of triple types and solution success</td>
<td>48</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Path Analysis of Mediator Effect</td>
<td>52</td>
</tr>
<tr>
<td>3.4</td>
<td>General Discussion</td>
<td>52</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>Experiment 2: Investigating Information Quantity Theory</td>
<td>56</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>56</td>
</tr>
<tr>
<td>4.2</td>
<td>Method</td>
<td>59</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Participants</td>
<td>59</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Design</td>
<td>59</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Materials and Procedure</td>
<td>59</td>
</tr>
<tr>
<td>4.3</td>
<td>Results and Discussion</td>
<td>60</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Task Success across Conditions</td>
<td>60</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Analysis of Triples Generated</td>
<td>61</td>
</tr>
<tr>
<td>4.3.2.4</td>
<td>Analysis of the Effect of Generating a Descending Triple</td>
<td>64</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Analysis of the Effect of Generating a Posvar Triple</td>
<td>66</td>
</tr>
<tr>
<td>4.4</td>
<td>Discussion</td>
<td>67</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>Experiment 3: The Goal Complementarity Theory of Dual Goal Facilitation</td>
<td>70</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>70</td>
</tr>
<tr>
<td>5.2</td>
<td>Method</td>
<td>72</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Participants</td>
<td>72</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Design</td>
<td>72</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Materials and Procedure</td>
<td>74</td>
</tr>
<tr>
<td>5.3</td>
<td>Results and Discussion</td>
<td>74</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Task Success across Conditions</td>
<td>74</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Triple Types across Condition</td>
<td>75</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Triple Generation Strategies Associated with Task Success</td>
<td>77</td>
</tr>
<tr>
<td>5.3.4</td>
<td>The Role of Descending Triples in Mediating Task Success</td>
<td>78</td>
</tr>
<tr>
<td>5.3.5</td>
<td>Discriminating between the Goal Complementarity and Contrast Class Information Accounts</td>
<td>80</td>
</tr>
<tr>
<td>5.3.6</td>
<td>Path Analysis</td>
<td>80</td>
</tr>
<tr>
<td>5.4</td>
<td>General Discussion</td>
<td>81</td>
</tr>
</tbody>
</table>
Chapter 6 Experiment 4: Extending the Contrast Class Theory of Dual Goal Facilitation to Different DAX Rules

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction</td>
<td>90</td>
</tr>
<tr>
<td>6.2 Method</td>
<td>92</td>
</tr>
<tr>
<td>6.2.1 Participants</td>
<td>92</td>
</tr>
<tr>
<td>6.2.2 Design</td>
<td>92</td>
</tr>
<tr>
<td>6.2.3 Materials and Procedure</td>
<td>92</td>
</tr>
<tr>
<td>6.3 Results</td>
<td>93</td>
</tr>
<tr>
<td>6.3.1 Task Success across Condition</td>
<td>93</td>
</tr>
<tr>
<td>6.3.2 Triple Generation Profile of Successful versus Non-Successful Participants</td>
<td>94</td>
</tr>
<tr>
<td>6.3.3 Triple Types across Condition</td>
<td>96</td>
</tr>
<tr>
<td>6.3.4 Discriminating between the Triple Heterogeneity and Contrast Class Accounts</td>
<td>97</td>
</tr>
<tr>
<td>6.4 Discussion</td>
<td>98</td>
</tr>
</tbody>
</table>

Chapter 7 Experiment 5: Testing the Contrast Class Account

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1 Introduction</td>
<td>100</td>
</tr>
<tr>
<td>7.2 Method</td>
<td>102</td>
</tr>
<tr>
<td>7.2.1 Participants</td>
<td>102</td>
</tr>
<tr>
<td>7.2.2 Design</td>
<td>102</td>
</tr>
<tr>
<td>7.2.3 Procedure</td>
<td>102</td>
</tr>
<tr>
<td>7.3 Results</td>
<td>103</td>
</tr>
<tr>
<td>7.3.1 Solution Success across Conditions</td>
<td>103</td>
</tr>
<tr>
<td>7.3.2 Production of Descending Triples</td>
<td>103</td>
</tr>
<tr>
<td>7.3.3 Analysis of Triple Types Produced across Condition</td>
<td>104</td>
</tr>
<tr>
<td>7.3.4 Analysis of Triple Types Produced across Success</td>
<td>106</td>
</tr>
<tr>
<td>7.4 Discussion</td>
<td>106</td>
</tr>
</tbody>
</table>

Chapter 8 Experiment 6: Manipulating Contrast Class Information across More than One Axis

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1 Introduction</td>
<td>112</td>
</tr>
<tr>
<td>8.2 Method</td>
<td>113</td>
</tr>
<tr>
<td>8.3 Results and discussion</td>
<td>113</td>
</tr>
<tr>
<td>8.3.1 Success Rates across Conditions</td>
<td>113</td>
</tr>
</tbody>
</table>
8.3.2 Production of Descending Triples ................................................ 113
8.3.3 Analysis of Triple Types across Condition .................................. 114
8.3.4 Analysis of Triple Types across Success ...................................... 115
8.4 General discussion ................................................................................ 116

Chapter 9 Experiment 7: Exploring Contrast Class Generation for
Presented Number Rules ............................................................................. 118

9.1 Introduction ........................................................................................... 118
9.2 Method .................................................................................................. 120
  9.2.1 Participants ................................................................................... 120
  9.2.2 Design .......................................................................................... 120
  9.2.3 Materials ....................................................................................... 120
  9.2.4 Procedure ...................................................................................... 121
9.3 Results ................................................................................................... 121
9.4 Discussion ............................................................................................. 122

Chapter 10.............................................................................................................. 124

10.1 Introduction ........................................................................................... 124
10.2 Overview of Experiments ..................................................................... 124
  10.2.1 Testing Evans' (1989) Positivity Bias Theory of DG Facilitation .... 124
  10.2.2 Testing Wharton, Cheng, and Wickens' (1993) Information Quantity Account of DG Facilitation 125
  10.2.3 Testing Wharton, Cheng, and Wickens' (1993) Goal Complementarity Theory ................................. 126
  10.2.4 Generalising the Contrast Class Theory to Other Rules..... 127
  10.2.5 Testing the Contrast Class Theory of DG Facilitation ...... 128
  10.2.6 Testing Contrast Class Theory by Manipulating Information across More than One Axis .......... 128
  10.2.7 Exploring the Generation of Contrast Class Information... 129
  10.2.8 Summary of Empirical Evidence ............................................. 129
10.3 Contrast Class Theory and the Extended Iterative Counterfactual Model ......................................................... 130
10.4 Evaluation of the Extended Iterative Counterfactual Model ................. 132
10.5 Testing the Extended Iterative Counterfactual Model ...................... 135
10.6 DG Facilitation and the Philosophy of Science ................................. 137
10.7 The "Hypothetical Thinking Framework" and Performance on the DG Task ................................................................. 141
10.8 Future Directions ............................................................................................................................. 143
10.9 Conclusion ........................................................................................................................................ 147

Chapter 11 References .......................................................................................................................... 149
List of Tables

Table 1-1 Theories of Dual Goal Facilitation, including Predictions and Empirical Tests ................................................................. 38

Table 2-1 Theories of Dual Goal Facilitation, including Predictions and Empirical Tests ........................................................................ 41

Table 3-1 Frequency of Correct Initial Rule Announcements (Solvers vs. Non solvers), Mean Numbers of Total Triples Produced, Types of Triples Produced (Posvars or Negative Types), and Triples Receiving Negative Feedback, by Condition .................................................................. 47

Table 3-2 Frequency of Correct Rule Announcements by Presence Versus Absence of Posvar Triples, and by Presence Versus Absence of Negative Triple ................................................................................................. 49

Table 3-3 Summary of Mediation Analysis of the Relationship Between Goal, Production of at Least One Descending Triple, and Success, Using Linear Regression (Including a Comparison of the p Values Obtained Using Linear and Logistic Regression) ........................................................................ 51

Table 4-1 Theories of Dual Goal Facilitation Examined in the Present Study, Including Predictions and Empirical Tests .................................................................................................................. 58

Table 4-2 Percentage of Successful Participants as a Function of Instruction Type and Number of Triples Generated ........................................................... 60

Table 4-3 Proportions of Triple Types Produced across Solvers and Non-Solvers ............................................................................................... 61

Table 4-4 Comparison of Proportions of Triple Types Generated across Triple Number Requirements ................................................................................................. 63

Table 4-5 Comparison of Proportion of Triple Types Generated across Instruction Type .............................................................................................. 64

Table 4-6 Frequency of Correct Rule Announcement by Presence or
Absence of Descending Triple ................................................................. 65

Table 4-7 Frequency of Descending Triple Generation as a Function of Instruction Type ................................................................. 65

Table 4-8 Frequency Counts of Presence of Descending Triple by Number of Triples Demanded .......................................................... 66

Table 4-9 Table 4.9 Frequency of Correct Rule Announcement by Presence or Absence of Posvar Triple and Presence of Posvar Triples by Instruction type .................................................................................. 66

Table 4-10 Frequency of Presence of Posvar Triple as a Function of Number of Triples Demanded ....................................................... 67

Table 5-1 Theories of Dual Goal Facilitation, Including Predictions and Empirical Tests ........................................................................ 73

Table 5-2 Frequency of Correct and Incorrect DAX Rule Announcements across Conditions ........................................................................ 75

Table 5-3 Participants' Triple-Generation Profile across Conditions ....................................................................................................... 76

Table 5-4 Triple-Generation Profile across Solvers and Non-Solvers when Collapsed across Conditions ................................................................ 77

Table 5-5 Frequency of Participants Producing at Least one Descending Triple across Conditions ................................................................. 79

Table 5-6 Presence versus Absence of at Least One Triple that was neither Descending nor Ascending across Conditions ............................. 79

Table 5-7 Summary of Mediator Analysis of the Relationship between Goal, Production of at Least One Descending Triple, and Success, Using Linear Regression ................................................................................ 81

Table 5-8 Definitions of DAX and MED Rules across Experiment 3 and the Present Experiment .................................................................. 84

Table 5-9 Percentage of Solvers across Rule and Success ....................................................................................................................... 85
Table 6-1 Frequency of Solvers and Non-Solvers across Conditions, Broken Down According to the Presence versus Absence of at Least One Contrast Class Triple ............................................................... 94

Table 6-2 Frequency of Solvers and Non-Solvers across Conditions, Broken Down According to the Presence versus Absence of at Least One Triple Containing 2 Equal Numbers ........................................................................... 95

Table 6-3 Triple generation profiles as a function of success ................................... 96

Table 6-4 Triple generation profiles as a function of instruction type ...................... 97

Table 7-1 Theories of Dual Goal Facilitation, Including Predictions and Empirical Tests ............................................................................................. 101

Table 7-2 Percentage of Correct and Incorrect DAX Rule Announcements across Conditions ................................................................................................. 103

Table 7-3 Percentage of Correct DAX Announcements by Presence versus Absence of at Least One Descending Triple .............................................................................. 104

Table 7-4 Percentage of Participants Producing at Least One Descending Triple by Condition ........................................................................................... 104

Table 7-5 Participants' Triple-Generation Profile across Conditions ..................... 105

Table 7-6 Participants' Triple-Generation Profile across Success ....................... 106

Table 8-1 Percentage of Correct and Incorrect DAX Rule Announcements across Conditions ................................................................................................. 113

Table 8-2 Frequency of Correct DAX Announcements by Presence vs. Absence of at Least One Descending Triple in Experiment 6 .................................. 114

Table 8-3 Summary of Triple Generation Profiles across Conditions......................... 115

Table 8-4 Summary of Triple Generation Profiles across Success on the Task ....... 116
List of Figures

Figure 5-1 Path Diagram Showing the Mediating Effect of Producing at Least One Descending Triple on Task Success ................................................................. 81

Figure 5-2 Iterative Counterfactual Model (Oaksford & Chater, 1994) ............... 87

Figure 7-1 Extended Iterative Counterfactual Model ........................................... 110

Figure 10-1 Extended Iterative Counterfactual Model for Non-Complementary Rules ................................................................................................................. 134
Chapter 1
Introduction to Hypothesis Testing Behaviour

1.1 Introduction

Hypothesis testing has been described as comparing internal thoughts to external facts in order to interact with the world (Poletiek, 2001). Evidence from day to day life suggests that humans are extraordinarily efficient at hypothesis testing; for example, language learning can be viewed as hypothesis testing, as can decision making about future possibilities (Evans, 2007). Evidence from empirical studies is, however, less flattering, with reported success rates in hypothesis testing tasks being very poor (e.g., Wason, 1960, Mynatt, Doherty, & Tweney 1978). Thus the primary purpose of this literature review is to explore the issues around hypothesis testing behaviour with a view to informing a programme of study which will allow exploration of this mismatch of real life evidence and empirical studies of hypothesis testing behaviours.

Formal hypothesis testing is seen as a stage in scientific discovery, and for this reason the literature surrounding the philosophy of science is initially reviewed, with particular attention being given to what has become known as the "problem of induction", that is the idea that future events will be similar to previous events, even though such a stance is logically erroneous. The problem of induction presents a challenge for the progress of science in that it involves making predictions about future events using evidence from previous events. Two philosophical arguments concerning the problem of induction are outlined: the Popperian (1959) approach of falsification, and the Bayesian probabilistic approach (Howson & Urbach, 2006), and differences between the two approaches are highlighted.

The nature of hypothesis testing behaviour is then examined in both real life settings (e.g. Mitroff, 1974:Tweney, 1985b) and in the laboratory across a range of hypothesis testing tasks (e.g. Bruner, Goodnow, & Austin, 1956; Wason, 1966). Comparing hypothesis testing behaviours in the real world to those observed in the laboratory allows some judgement to be made about the ecological validity of the necessarily simplified tasks that are undertaken in empirical studies. A key issue, for example, is whether the strategies employed by lay people are similar to or different from those deployed by trained scientists. In this section descriptions of hypothesis
testing behaviours are presented; for example, the concept of simple enumeration is introduced (i.e., the idea that examples thought to belong to a target class are tested), which is then related to the concept of confirmation bias (i.e., the tendency to confirm rather than to falsify hypotheses). The detrimental effect of confirmation bias on successful hypothesis testing (and some of the controversy surrounding the definition of confirmation bias) is explored, and studies which attempt to overcome the bias are presented. Other theoretical ideas concerning hypothesis testing behaviours that are introduced include the notion that people apply a positive test strategy (Klayman & Ha, 1987, 1989), and that people make considerable use of counterfactuals (Farris and Revlin 1989a, 1989b).

The review then moves on to examine the literature surrounding the standard paradigm for hypothesis testing behaviours, that is, Wason's (1960) 2-4-6 task. The task is described in both the standard Single Goal (SG) form and its Dual Goal (DG) variant which is logically identical but has very different patterns of performance with much higher success rates being reported in the DG form of the task. The argument is made that a successful account of the DG facilitatory effect would go some way to explaining poor performance on the standard SG task, and thus allow some insight into hypothesis testing behaviours in the laboratory. To this end theoretical explanations of the facilitatory effect of DG instructions are examined alongside the evidence which has been presented for these accounts, and the specific predictions from each theory. These predictions rely to a large extent on examining quantitative and qualitative changes in the types of tests that are designed to test local hypotheses, thus a brief description of the coding of triples (which are used to test hypotheses in the task) is provided. Finally a table summarising the theories, predictions and empirical tests of the theoretical accounts is presented which provides a framework for the programme of study in this thesis.

1.2 Review of the Literature

1.2.1 Scientific Reasoning and the “Problem of Induction”

It has been claimed that scientific reasoning is of interest to psychologists in that it is analogous to humans trying to acquire knowledge about their everyday environment (cf. Poletiek, 2001), a view which echoes that expressed by Kelly (1955), who likened the acquisition of knowledge to scientific reasoning. Einstein (1950) also took the view that equivalent thought processes underpin scientific and everyday reasoning, albeit with scientific reasoning demanding more precise definitions of concepts and conclusions alongside more careful design of experiments. Some
philosophers of science have also agreed with claims for parallels between scientific and everyday reasoning, noting that the acquisition of scientific knowledge is a special case of everyday reasoning in that it is more systematic in its approach (Popper, 1989).

The importance of the role of science to how philosophers view the discovery of knowledge in the everyday world is illustrated by Locke acting as a type of philosophical under-labourer to Newton in the belief this would lead to him discovering how we come to know anything and how we understand the fundamental structure of the world (Couvalis, 1997). The success of Newtonian physics in making verifiable predictions led other thinkers to believe this strategy to be the correct course of action, to such an extent that philosophers during the Enlightenment came to believe that science would provide the underpinnings of the correct social system and in this way replace metaphysics, ethics and political philosophy (Couvalis. 1997).

This latter view, however, was largely undermined by Hume (Proctor & Capaldi, 2006) who outlined what has become known as the “problem of induction”. He noted that whenever we make inductive inferences (i.e., reason from particular instances to a general law), we presuppose what he termed the “uniformity of nature” (UN); that is, we assume that instances of a phenomenon not yet examined will be similar in relevant respects to those previously examined. Hume went on to question how such an assumption could be justified. He argued that it is easy to imagine a universe that is not uniform but which changes its course from day to day, and since such a “non-uniform” world is conceivable then we cannot prove the truth of UN because such a proof would then show the existence of a non-uniform world to be a logical impossibility. Hume further argued that not only would it be impossible to prove the existence of UN, it would also be impossible to provide good empirical evidence for its existence: while it is a fact that nature has behaved in a uniform way thus far, we cannot appeal to UN as evidence that it will continue to so in the future, as this would be appealing to uniformity of nature as evidence for its own existence.

This intriguing argument has exerted, and continues to exert, a powerful influence on the philosophy of science, not least since science (and therefore modern technological society) appears to rely on induction, and Hume’s argument indicates that induction cannot be rationally justified. There have been a variety of responses to the problem of induction; for example the use of probability to suggest that although the premises of an inductive problem do not guarantee a logically valid conclusion, the evidence from previous experience shows that the conclusion is highly probable. It should be noted, however, that this then leads to the question of what constitutes a high probability. A second response is that induction is so fundamental to thinking
and reasoning that it is not the sort of thing which can be justified, and that to demand such a justification is therefore incoherent. Strawson (1952) defended this view, which dissolves rather than solves the problem with the following analogy: if a person is concerned about the legality of an action they would consult a law text and compare the action with the text. However, suppose someone was worried about law itself, how could they check the legality of the law? The law is the standard against which legality is judged and it therefore makes little sense to question whether the standard itself is legal. Strawson suggests that the same argument can be made for induction. He claims that induction is one of the standards used to decide if claims about the world are justified and it thus makes no more sense to question if induction is justified than it does to question if the law is legal.

Popper’s (1972) response to the problem of induction was to argue that the problem does not in fact exist: we do not discover irregularities but rather conjecture that such irregularities exist. Science can then be said to be rational as we can sometimes falsify conjectures by confronting them with basic statements with which they are inconsistent. If we are lucky we will stumble on true conjectures; however, even if we do not we will learn from the falsification of our conjectures. Popper thus argued that as falsification is logically deductive there is no need for a logic of confirmation; indeed, there is no logic of confirmation. Popper distinguishes two sides to Hume’s problem of induction: (a) the logical problem (i.e., are we justified in believing that unexperienced instances of certain phenomena will be similar to experienced ones?); and (b) the psychological problem (why do reasonable people believe that unexperienced instances will be like experienced instances?).

Popper claimed to have solved the logical problem by showing that there is no role for induction in science. If we assume that there are certain regularities in the world, and that these postulated regularities can be falsified by finding counterexamples, then this, according to Popperian philosophy is all that is required for science. Thus, scientific reasoning can be reconstructed as being deductive. An example of this is that if we have a conjecture that all swans are white then a single example of a black swan will refute that conjecture. Popper’s solution to the psychological problem of induction is that we conjecture the existence of certain regularities and test our conjectures. If our conjectures are not refuted then we temporarily accept them, and there is thus no need to assume the existence of an inductive procedure.

One tenet of Popper’s approach to scientific reasoning is that science best progresses by making bold conjectures which deal with largely unobserved phenomena. This view is tied to a second point: that when testing theories scientists
Introduction to Hypothesis Testing Behaviour

should use the notion of corroboration rather than confirmation. Corroboration is a comparative notion; that is, it has no absolute value. A theory is said to be more corroborated than a rival if it is more falsifiable, and has undergone more rigorous testing but has not yet been refuted (Popper, 1989). Thus, in terms of Popper’s philosophy, a theory will gain most support if subjected to severe testing; that is sincere attempts have been made to falsify it. A recent challenge to this view comes from Poletiek (2001), who notes the more severe the test the less its potential to falsify the hypothesis. Thus, if a severe test is chosen the outcome will be either weak refuting evidence (because of the severe test) or strong supporting evidence; therefore a severe test is only really informative if it supports the hypothesis, as a refutation using a strong test is only weak evidence against the theory.

An implicit assumption of the Popperian approach to hypothesis testing is that a hypothesis under consideration is seen as representing “the state of the world” until a disconfirming instance demonstrates this not to be the case: no account is taken of the plausibility of the hypothesis or the scientist’s confidence in its veracity. At an intuitive level this seems to be descriptively weak: it is likely that on initially forming a hypothesis a scientist’s confidence in it may be tentative, and that confidence (and therefore belief) in the hypothesis will vary as a function of evidence relating to it. In contrast to this “all or nothing” approach the Bayesian account suggests that belief in a hypothesis is adjusted in the light of empirical evidence. Initially proposed as a theory of subjective probability and belief revision, Bayes’ Theorem has been developed into a philosophy of science (Howson & Urbach, 2006). According to this approach, probability is seen as a degree of belief, thus belief in a hypothesis (and therefore the subjective probability of its veracity) varies as a function of prior belief in a hypothesis and the diagnosticity of the evidence, formally expressed in the case of two competing hypotheses as follows:

\[
\frac{P(H_1 | D)}{P(H_2 | D)} = \frac{P(H_1)}{P(H_2)} \times \frac{P(D | H_1)}{P(D | H_2)}
\]

in which \(H_1\) is a focal hypothesis, \(H_2\) is the competing hypothesis, \(D\) is the diagnosticity of the evidence and \(P\) is probability. This can be expressed more informally as follows:

Posterior belief = Prior odds X Likelihood ratio

It therefore follows that iterative testing leads to gradual revisions of belief in hypotheses as evidence either supporting or appearing to refute them is advanced. It is of note here that in his overview of the Popperian and Bayesian approaches to the
problem of induction Evans (2007) states that in the case of two competing, exhaustive hypotheses a good test of a hypothesis is equivalent from a Popperian or Bayesian perspective.

The problem of induction notwithstanding, it is clear that science rests on the production and testing of theories. Testing of theories is by a cyclical process of activity: hypotheses are derived from a theory, predictions derived from these hypotheses, and these predictions are then tested (Poletiek, 2001). Empirical testing either supports or fails to support the predictions made, with the effect of supporting or failing to support the theory from which they are derived. If the empirical evidence supports the hypotheses then the theory can be sustained. On the other hand, if the empirical evidence either refutes or does not fully support the hypotheses then the underlying theory is examined and either rejected or amended to accommodate the new information. This process is known as the “scientific method” (Horst, 1931). Hypothesis testing therefore consists of two distinct phases: the production of a hypothesis and the testing of the hypothesis. More formally, these phases are designated as the “context of discovery” and “the context of justification” (Reichenbach, 1938). Hypothesis testing behaviour forms the main focus of the present thesis.

1.2.2 The Nature of Hypothesis Testing Behaviour

Early work on hypothesis testing behaviour investigated hypothesis generation and testing of participants using a categorisation task (e.g., Bruner et al., 1956). Bruner et al. considered this type of cognitive activity to be ubiquitous and a prerequisite of all other cognitive tasks. In one of their studies Bruner et al. presented participants with an array of 81 instances of a stimulus which varied on four attributes, with three possible values of each of these attributes, that is: shape (square, circle, cross), number of figures (one, two, three), colour (red, green, black) and number of borders (one, two, three). Thus all possible combinations of the attributes were available. Participants were required to discover the researcher’s “concept” defined in terms of these attributes, by successive testing of individual stimuli in order to identify whether they exemplified the concept.

Bruner et al. were interested in the strategies used by participants to discover the researcher’s concept. One strategy which they identified was “enumeration”, that is, testing hypotheses derived from information gained by positive examples of the concept. However, while such a strategy can lead to correct identification of a sufficient hypothesis, it cannot identify whether such a hypothesis is both necessary and sufficient. Thus, should the target concept be more general than the participant’s
Chapter I

Introduction to Hypothesis Testing Behaviour

hypothesis, an enumerative strategy would not identify the restricted nature of their hypothesis. Take, for example, the true concept “red shapes”. If the concept “red squares” is formed and only shapes which display both of these attributes are selected, then the true nature of the concept will never be discovered. As this example illustrates, it is logically possible to arrive at an incorrect concept by enumeration because only positive evidence is utilised. Note, however, that a single negative test of the concept (i.e., selection of a “non-square” shape) leads to the identification of the overly-restricted nature of the hypothesised concept.

Whilst the work of Bruner is usually characterised as being concerned with concept formation (e.g., Wason, 1960), Poletiek (2001) neatly summarises the relationship between inductive reasoning and hypothesis testing behaviour when she observes that Bruner et al.’s participants must have produced abstract hypotheses by categorising old information which had been tested against previous experience, and then processed and produced new information based on these abstract hypotheses. In this manner inferences are made about the categories, causing generalisation of information contained within the categories and leading to the testing of the boundaries of these categories by selecting examples which were assumed to fit into the categories, thus extending and defining categories by a process of hypotheses generation and testing.

Poletiek (2001) describes several stages in hypothesis testing. In a given situation the reasoner must initially produce a hypothesis about the current situation and then devise a test of this hypothesis, the outcome of which must provide information about the truth of the hypothesis. The test is performed and the outcome observed. The information gained must then be assimilated into the old information and the impact on the hypothesis assessed. This cycle of activity is akin to the scientific method in which similar steps are followed. One assumption of the cycle of hypothesis testing described here is that the reasoner seeks new information, that is, they seek to extend the body of evidence on which to evaluate their hypotheses. A question which thus arises is how the reasoner chooses their test; it is evident that given any situation the realm of possible tests is likely to be extensive, and that the reasoner is constrained in the number of tests to be performed by the demands of the current situation. A reasonable assumption would be that the reasoner would choose the most informative test, however, this in itself is likely to be problematic in that the reasoner does not know either the truth status of their working hypothesis, or its relationship to the truth. As noted above, Popper (1959) advocated the use of falsification in hypothesis testing. This strategy is based on the observation that any number of confirming instances are unable to verify the truth of a hypothesis, whereas
a single disconfirming instance will falsify it. Thus, according to Popperian ideals, the
greatest support for a hypothesis can be gained from a “strong” test, (i.e., one which is
most likely to falsify the hypothesis) but which in fact does not do so and, therefore,
“good” scientists will follow a strategy of falsification when testing their ideas.

Empirical studies of the hypothesis testing strategies of scientists have, however,
demonstrated that despite the tenets of Popperianism, scientists may well tend to
adopt a confirmation strategy. For example Mahoney and Kimper (1976) used the
Wason (1966) selection task to examine the hypothesis testing strategies of a group of
scientists. In this task participants are shown four cards each of which has a number
on one side and a letter on the other side. The facing sides of the four presented cards
show respectively a vowel, a consonant, an even number and an odd number.
Participants are asked to select which cards are necessary to verify the truth of the
statement: “If there is a vowel on one side of the card, then there is an even number on
the other side”. To verify the rule participants should select the vowel (to determine if
there is an even number on the obverse) and the odd number (to ensure that there isn’t
a vowel on the back, which would falsify the rule). The remaining cards are
uninformative as to the truth status of the rule. Using a sample of physicists,
biologists, psychologists and sociologists, Mahoney and Kimber found that only 10%
of participants selected the logically necessary cards. Similarly, Kern, Mirels, and
Hinshaw (1983) in their investigation of the performance by scientists on the task
found that just 20% identified the logically correct cards.

Whilst these latter success rates on the task are somewhat superior to the 4%
reported when using “ordinary” participants (Wason & Johnson-Laird, 1970), they
still suggest that scientists may be prone to a confirmation bias. There is, however,
some evidence of such results being an artefact of the instructions used (Tweney &
Yachanin, 1985). In their study involving scientists from a range of disciplines
Tweney and Yachanin found that when asked to “find out if a rule is true or false” as
opposed to “testing” it, then 90% of their scientists selected the falsifying cards.
Tweney and Yachanin argued that this instructional artefact was due to the limited
experience that scientists have with published research reporting negative results. It is
more common practice in scientific papers that hypotheses are proposed and then
tested by presenting empirical results which report positive effects. As a result, the
term “test” suggests to scientists a confirmatory heuristic. According to this view the
apparent confirmation bias displayed by scientists may be the result of cultural
interpretation of the term “test” rather than a bias per se. This observation cannot,
however, be extended to explain poor levels of performance among lay participants;
indeed, it would be interesting to apply this manipulation to a lay population to investigate if these same patterns of performance are evident.

More recently the interpretation of the card choice in the standard selection task as representing a confirmation bias has become less pervasive, with poor performance being attributed to a “matching bias” in which card choices tend to mirror the lexical content of the conditional statement which is to be tested (Evans, 2007). Early evidence of this came from a study by Evans and Lynch (1973) which used the negations paradigm (e.g. “If there is a vowel on one side of the card, then there is NOT an even number of the other side”). By systematically varying the negated component of the conditional statement to be tested Evans and Lynch demonstrated both the tendency to match the terms in the conditional statement when making logically correct card choices, and also that when matching was controlled for the true consequent (vowel) and false consequent (not a vowel) card were equally preferred. Despite these apparently clear findings other studies have presented evidence that suggests patterns of performance may be attributable to a combination of matching and confirmation biased responses (Krauth, 1982; Reich & Ruth, 1982). More recent evidence has shown that although matching effects are robust when negation is implicit, the effect is reduced when the negation is made explicit (Stahl, Klauer, & Erdfelder, 2008). As such, theoretical accounts of performance on the task remain somewhat unclear. It should be noted, however, that studies employing the negations paradigm have used lay participants so it is somewhat difficult to draw conclusions as to the reasoning processes of scientists from them. It is however noteworthy that in a study of the selection task which used scientists as participants both the patterns of card choice and the post hoc rationalisations of those choices when using both abstract and thematic content were similar to those reported in studies using lay populations (Griggs & Ransdell, 1986), providing some evidence that it would be reasonable to generalise from lay to scientific populations.

In discussing evidence for the selection task as relating to hypothesis testing it is important to note the differing demands of the selection task and the 2-4-6 task, and therefore the resulting impact on hypothesis testing behaviour. In the selection task the reasoner is given a hypothesis and asked to test it in a single, all-or-nothing experiment. Traditionally, therefore, the selection task has been characterised as Popperian in the demands that are made of the participant (although see Oaksford & Chater, 1994, for a probabilistic theory of task performance). No account is taken of the reasoner’s belief in the hypothesis and evidence is not collected by the reasoner which allows them to vary that belief. By contrast, in the 2-4-6 task (introduced in the next section) the reasoner generates their own hypothesis, and is expected to collect
evidence concerning it; thus their subjective belief in their current hypothesis may vary both as a function of their initial confidence in their proposed hypothesis and of the evidence they obtain. The demands of 2-4-6 task may therefore be characterised as requiring a more Bayesian approach to hypothesis testing. As such it is important to take care when drawing general conclusions from findings across the two tasks.

Evidence from "real life" studies of science also tends to indicate the existence of a confirmation bias in scientists. In his review of the work of Michael Faraday, Tweney (1985a) noted on analysing Faraday's laboratory books that out of a set of 135 experiments conducted in 1831 as part of his discovery of electromagnetism, Faraday appeared to follow a strategy of confirmation in the initial stages of the project; that is, he attempted to confirm his ideas, and only when he had gathered strong evidence as to the veracity of a theory did he move on to a falsificatory strategy. Similarly, Mitroff (1974) in his work with NASA lunar scientists noted that they appeared to employ a confirmation strategy in their work, as did Mahoney (1976) in his examination of the strategies employed by scientists in the real world. This evidence tends to suggest that contrary to the views of Popper, a tendency to confirm hypotheses may indeed be a useful strategy at least in the early stages of scientific research. Tweney and Chitwood (1995) suggest that in Faraday's case at least, this was because he was taking into account the fragility of his new hypotheses and was amassing sufficient confirmatory evidence before adjudging that making a falsifying test was worthwhile.

In their empirical investigation of modes of enquiry in hypothesis testing, Kareev, Halberstadt, and Shafir (1993) gave participants a binary categorisation task and allowed them either to receive information about possible exemplars of the categories or to generate their own examples. Results showed that in early stages of the task participants tended to prefer the reception mode, moving to a more active experimentation phase only after some data had been obtained. Kareev et al. interpret their findings as offering support for the notion that there are two modes of scientific enquiry: observation and experimentation. They characterise the early hypothesis testing behaviour of their participants as "observe and formulate", stating that previous empirical and theoretical work indicated that hypothesis generation is the outcome of generalisations based on positive cases. Given the lack of naturally occurring instances of scientific phenomena, it is possible that when engaged in a programme of study, scientists aim to emulate this base of positive cases by devising experiments which are likely to provide positive instances of their working hypothesis. It should be noted, however, that although there might be a paucity of naturally occurring evidence for a scientific theory, even in the early stages of
hypothesis testing it is a prerequisite that some evidence must exist for the theory to be formulated.

In their study of hypothesis testing in an artificial universe Mynatt, Doherty, and Tweney (1978) noted that the normative strategy of falsification was of little utility in their complex task, and that the most successful participants were those who adopted both confirmatory and disconfirmatory strategies. In their study of participants’ behaviours when asked to discover the function of a novel button on a computer controlled robot Klahr and Dunbar (1988) showed that when using naturalistic materials their participants were able to appreciate the benefits of disconfirming information when it was elicited, even though they tended to follow a positive test strategy (Klayman & Ha 1987, 1989) in which tests were expected to be cases of the hypothesis. Tweney, Doherty, and Mynatt (1981) stated that the confirmation bias was not a single entity, but that at least four different types were evident: (i) the failure to seek disconfirmatory evidence, (ii) the failure to utilise disconfirmatory evidence when available, (iii) the failure to test alternative hypotheses, and (iv) the failure to consider whether evidence for a working hypothesis may also support another hypothesis. As Tweney and Chitwood (1995) note, the latter two types of confirmation bias imply the existence of alternative hypotheses, whereas the first two refer only to a single hypothesis. This is of interest in that it implies that strategies employed in testing a single hypothesis might vary from those used when comparing two competing hypotheses.

In their work on the use of counterfactuals in hypothesis testing Farris and Revlin (1989a, 1989b) went some way to addressing the issue of the role of the testing of alternatives in successful hypothesis testing. Using a novel analysis of participants’ protocols in which they compared their current tests to all previous hypotheses, Farris and Revlin demonstrated that contrary to the contemporary view of people’s hypothesis testing strategy as being primarily confirmatory, many participants’ strategies could instead be characterised as “counterfactual”, with reasoners entertaining multiple hypotheses. This counterfactual strategy was, moreover, claimed to be the best single predictor of successful reasoning. In addition, Farris and Revlin showed that in their study successful reasoners could not be distinguished from unsuccessful ones in terms of their use of a disconfirmatory strategy. Farris and Revlin further noted that Faraday (in Tweney, 1985a) appeared to be employing a methodology akin to the counterfactual strategy as he juxtaposed competing hypotheses by combining his investigations of magnets and electromagnets and alternating between them.
In their research examining the use of counterfactuals in hypothesis testing, Farris and Revlin used what has become the standard paradigm for investigating this behaviour: Wason's (1960) rule discovery task. The following section describes the task in both its standard single goal (SG) format as well as its dual goal (DG) format which has featured extensively in more recent research. In addition, the following section reviews the patterns of performance displayed on these task variants and examines the theoretical implications of such performance patterns.

1.2.3 Wason’s (1960) 2-4-6 Rule Discovery Task

While the Wason selection task (1966, described in Section 1.2.2) provides a useful paradigm to examine hypothesis testing strategies, it fails to mimic true scientific and everyday reasoning in that participants are provided with a single hypothesis and a limited pool of possible tests to choose from. A more valid paradigm which mimics such reasoning more closely is the 2-4-6 task (Wason, 1960). Wason originally devised the 2-4-6 task to investigate whether people naturally adhered to Popper's (1959) norm of falsification. As noted earlier, this norm hinges on the principle that whilst a single disconfirming instance can reveal that a hypothesis is incorrect, any number of confirming instances can never prove that the hypothesis is true. In the standard 2-4-6 task participants have to discover a rule that generates sequences of three numbers (triples). Participants are initially given an example, conforming triple (“2-4-6”), and they then produce further triples that the experimenter classifies as conforming or not conforming to the rule. Participants generate triples until they are confident they know the rule, at which point they announce it. The to-be-discovered rule is “three ascending numbers”.

Wason (1960) described the 2-4-6 task as differing from Bruner et al.'s (1956) task in several respects. First, rather than categories being discovered, the properties of which need to be enumerated (e.g., red and square), the to-be-discovered rule in the 2-4-6 task is relational and therefore can be described in more abstract terms. Second, the universe of possible tests is infinite rather than being constrained by the possible examples offered in previous studies. The finite pool of possible tests previously used places a constraint on the scope of inductive thinking in that the pool of necessary and sufficient exemplars will, by definition, always be smaller than the pool of merely sufficient examples. Thus, the number of possible necessary and sufficient tests is limited allowing the assiduous participant to exhaust the pool of necessary and sufficient tests and thereby carry out a test of a merely sufficient example. Third, the instances are not presented as stimuli. Instead they have to be generated by the reasoner with the effect that an additional task demand becomes that of deciding on the type and number of tests to be performed. As Wason (1960) notes, in standard
concept categorisation tasks all possible instances are displayed and the reasoner is aware that all the evidence is available to them, giving them an opportunity to judge the scope of the required evidence collection. Finally, in the 2-4-6 task, unlike concept categorisation tasks, participants keep a written record of their tests and subsequent outcomes, thus removing any effect of memory constraints from the task.

Despite the apparent simplicity of the 2-4-6 task people perform poorly, with only around 20% announcing the correct rule at their first attempt (e.g., Tukey, 1986; Wason, 1960). Many incorrect announcements are restricted versions of the target rule, such as “numbers increasing by two” (Kareev et al., 1993), that is, participants tend to announce rules which are sufficient, but not necessary. Wason (1960) also noted that solvers and non-solvers could be differentiated in terms of both the number of triples produced, with solvers generating more, and the type of triples produced, with solvers generating a higher proportion of triples receiving negative feedback (cf. Klayman & Ha, 1989b). Wason interpreted the non-solvers’ strategy of testing positive instances of hypotheses as a failure to appreciate the benefits of falsification and therefore reflecting a confirmation bias – as discussed above in the context of other hypothesis testing paradigms (e.g., Gorman & Gorman, 1984; Mynatt, Doherty, & Tweney, 1977). Wason further cited evidence of this confirmation bias from participants’ protocols; he noted some participants, following incorrect rule announcements, would simply reword their hypothesis and then continue to test it and then announce a logically identical rule.

Efforts to improve performance by instructional manipulations emphasising falsification have generally proved unsuccessful (Gorman & Gorman, 1984; Gorman, Stafford, & Gorman, 1987; Kareev et al., 1993; Mynatt, Doherty, & Tweney, 1977; Mynatt et al., 1978; Tweney et al., 1980). For example, in a series of four experiments Tweney et al. (1980) attempted to determine whether instructions to falsify could improve performance on the task. Experiment 1 attempted to compare performance when direct instructions to falsify were employed. Participants in the confirmatory condition were given a hypothetical rule of “three equal numbers” and an example triple of 3-3-3. They were told they could test the hypothesis with positive examples of the hypothesis (such as 8-8-8) and if such a triple fitted the rule that would be evidence for the rule. By contrast participants in the disconfirmatory condition were given the same hypothesis and exemplar triple and told that if tested with a triple which didn’t fit their hypothesis triple such as 5-7-9, and that this triple did indeed fit the rule then they knew the hypothesis was wrong. While falsification instructions appeared to influence the test strategy employed, raising the number of negative triples generated (that is, triples which do not fit the rule) this did not translate into
improved performance. In Experiment 2, Tweney et al. (1980) explored the idea that disconfirmatory instructions would be more helpful to participants at a later stage of the study. Consequently they varied their methodology such that strategy instructions were given after the first rule announcement. As with Experiment 1, changes were observed in the types of triples produced, but again this did not translate into differential performance on the task across conditions.

In contrast to Tweney et al.’s null results, Gorman and colleagues (Gorman, Gorman, Latta, & Cunningham, 1984) compared success rates across confirmatory and disconfirmatory instructions using both groups and individuals as well as two different tasks. In Experiment 1 and 2 they used an alternative task, the New Eleusis problem, which asks participants to discover the rule for laying out playing cards. They showed that instructional manipulations could indeed affect both the types of strategies employed by participants and also success rates, although they note that use of disconfirmation was limited to approximately 50% of tests. In their final study they generalised the disconfirmatory instruction they were using to the 2-4-6 task and again showed facilitated performance. Gorman (1992a) attributed the improved performance in his studies to the fact that participants were not given feedback on putative rule announcements, but were instead informed that it was up to them to decide if their rule was correct. Gorman suggests that by removing this second (and more powerful) source for testing hypotheses, participants could no longer rely on the experimenter for confirmation or disconfirmation, and that confirmation might be an effective heuristic when participants can no longer appeal to an outside authority for information.

Gorman (1995) suggests that one reason for the general failure of disconfirmatory instructions to improve performance might be related to a misconception of the term “confirmation”. He notes that in an earlier study (Gorman, 1992a) he had used what he termed “confirmatory” and “disconfirmatory” instructions, but that these instructions asked participants to try to produce triples which were “right” and “wrong” respectively. This goes to the heart of one of the early criticisms of the confirmation bias approach. Wetherick (1962) was an early critic of the 2-4-6 task noting both that the exemplar triple induced the reasoner to produce an overly-restricted rule and that Wason’s definition of an eliminative triple (one which was not compatible with a previously stated reason for a choice) was erroneous. Wetherick argued that a triple was only falsifying when it was either: (i) an opposite example of a hypothesis for which positive feedback was expected (i.e., the participant assumed that they would be told that the triple conformed to the experimenter’s rule), or (ii) an example of a hypothesis for which negative feedback
was expected. In a study designed to test this idea Wetherick asked participants to indicate whether they believed their test triple would or would not conform to the rule. In this way he was able to code accurately the participant's intention for each test. Wetherick found that falsificatory tests were relatively rare, with less than 10% of triples produced being disconfirmatory according to his revised definition.

Poletiek (2001) agreed with this view of disconfirmation, citing evidence from an earlier study of her own (Poletiek, 1996b) in which participants who were asked to employ a falsification strategy preferred negative tests but also expected these tests to fail. However, Poletiek argued that the situation is even more complex in that participants are in fact unable to produce falsifying tests. Interestingly this view of disconfirmation is directly contrary to that of Tweney et al. (1980); in their study they specified a "confirming" test as one which conforms to the experimenter's rule and a disconfirming test as one which was not expected to conform to the rule. Whilst the basic premise of Wetherick and Poletiek is appealing (i.e., that a disconfirming triple must be either a positive example of the participant's rule for which negative feedback is expected, or a negative example of the participant's rule for which positive feedback is expected) these complex definitional issues will be sidestepped in the present thesis by simply referring to positive tests as ones which conform to the experimenter's rule, and negative tests as ones which do not conform to the experimenter's rule.

It remains moot why performance fails to improve in the face of disconfirming evidence, although the emphasis of these studies on manipulating the so-called "context of justification" (i.e., hypothesis testing) rather than the "context of discovery" (i.e., hypothesis generation) may well be implicated. It could be argued that both contexts are important for success on Wason's rule discovery task. Indeed, it may be that without an alternative hypothesis in mind reasoners are simply unable to make use of disconfirming information.

Evans (2007) offers an alternative explanation for poor performance on the task which is based within his "hypothetical thinking framework" (a revision and extension of his heuristic-analytic model which attributed cognitive biases to selective processing; see Evans, 1983; 1989) which he situates within a broader "dual process" view of reasoning. Dual process theory, as the name implies, centres on the notion that there are two distinct processes involved in cognitive tasks. Dual process theory is itself a weaker version of the Dual System account of cognition which proposes that the two processes have differing evolutionary histories and originate in different neural substrates (Evans & Over, 1996; Stanovich, 2004). Evans (2006) notes that this stronger account is of necessity broad, and as such may not apply to
specific cognitive tasks such as hypothesis testing. For this reason he situates his hypothetical thinking framework at an intermediate level of abstraction which allows models of specific tasks (including Wason's, (1960), rule discovery task) to be developed. Evans (2006) identifies the following as attributes of heuristic processing:

Unconscious and automatic
Rapid, powerful and parallel processing
Associative
Not requiring working memory resources
Not capable of domain general processing
Low effort

By contrast, analytic processing is seen as:

Responsive to verbal instructions (Evans & Over, 1996; Stanovich, 1999)
Using slow and sequential processing
Linked to general cognitive ability
Capable of domain general reasoning
Volitional and controlled

Evans' (2007) hypothetical thinking theory consists of three parts: (i) a set of three principles that aim to describe the general nature of hypothetical thought (Evans, Over, & Handley, 2003), (ii) a proposal that thinking involves the manipulation of mental representations in the form of epistemic mental models, and (iii) a processing model. The first of the three principles is the singularity principle, the notion that people consider only one hypothetical possibility or mental model at a time. Evidence for this idea comes from the work of Mynatt, Doherty, and Dragan (1993) who showed that in a pseudodiagnosticity task people were only able to maintain and operate on one hypothesis at a time in working memory, and from poor performance levels in the THOG problem (Wason & Brooks, 1979), a disjunction task described by Evans (2006) as one of the most difficult reasoning tasks examined in the laboratory, which requires participants to hold two hypotheses simultaneously. The relevance principle relates to the idea that people only consider the most plausible alternative in the current context, and the satisficing principle suggests that models are evaluated according to current goals and accepted if satisfactory. Satisficing is seen as contrasting with optimising; whereas optimising is finding the best possible solution, satisficing is settling for that which is “good enough”.

- 28 -
Evans (2007) suggests that it is the combination of the singularity and the satisficing principle which leads to suboptimal performance on hypothetical thinking tasks (and by extension the 2-4-6 task). The singularity principle will lead to the consideration of a single hypothesis which is then evaluated according to the satisficing principle; if this first hypothesis is found to have satisfied the demands of the task then it is accepted without consideration of further alternatives. Note that both of these principles are concerned with the context of justification, and whilst Evans does not make it explicit it is highly likely that the relevance principle is involved in the generation of the initial, overly-restricted hypothesis. Indeed, recent work by Cherubini, Castelvecchio, and Cherubini (2005) has shown that when formulating hypotheses participants retain the maximum amount of relational information available in the seed triple; it could, therefore, be assumed that the participant will establish an initial hypothesis (H) of “even numbers ascending by 2”. In another interesting piece of recent research, Van der Henst, Rossi, and Schroyens (2002) demonstrated that when participants believe that the exemplar triple is generated by chance then success rates on the task rise substantially, indicating that participants believe that the information contained in the initial exemplar is highly relevant to the rule.

The processing model for Evans’s 2007) hypothetical thinking theory assumes that the two systems compete for control of behaviour and that by default the systems therefore work in parallel. Evans proposes that heuristic processes set up default responses which occur unless analytic processes intervene, although he notes that analytic processes are always involved if only to “approve” the default response produced by heuristic processes. The hypothetical thinking framework can thus lead to bias by two main mechanisms: heuristic processing may produce bias through selective representation and contextualisation (Evans 1989; Stanovich, 1999); alternatively, the singularity and satisficing principles may lead to bias via analytic processing. In his review of performance on the 2-4-6 task Evans (2007) tentatively attributes the positive test bias (i.e., people’s tendency to test positive instances of their current hypothesis) to a heuristic relevance effect, but further suggests that the analytic system has a role in maintaining an erroneous hypothesis despite negative feedback following a rule announcement due to its role in rationalisation and confabulation.

An alternative account of the hypothesis testing strategy adopted by participants in the 2-4-6 task is that of the Iterative Counterfactual Model (ICM), proposed by Oaksford and Chater (1994). The ICM builds on the work of Farris and Revlin, (1989a, 1989b) but addresses what they describe as a logical shortcoming in this
account. According to the Farris and Revlin (1989a) model, a participant generates a hypothesis on the basis of the seed triple, for example, “even numbers”. The complement of this hypothesis (e.g., “odd numbers”) is then generated and tested using a positive test strategy (Klayman & Ha, 1987, 1989). If “yes” feedback is given then the even numbers hypothesis is seen to be falsified and the odd numbers hypothesis is seen as plausible. However, as Oaksford and Chater point out, this is logically inconsistent because the complementary “odd numbers” rule is falsified as it is generated; thus discovering that the triple 3-5-7 is an instance of the rule (in addition to the original 2-4-6 exemplar) indicates that both the even numbers and the odd numbers rules must be incorrect. Farris and Revlin (1989b) revised their model to suggest that the counterfactual strategy is only used to generate falsifying instances of a hypothesised rule. In the revised account “yes” feedback to an instance of a complementary hypothesis (H’) leads the participant to reject H’ as the expanded set is not large enough. Oaksford and Chater note that while it is correct to reject H’ the reason given is logically incorrect, as is Farris and Revlin’s proposal that at this stage H is not rejected, rather that a new counterfactual hypothesis is generated. Oaksford and Chater address these shortcomings by suggesting that on receiving “yes” feedback to an instance of H’, this triple is added to the list of instances of the original rule and a new hypothesis is generated that is based on examining the list of instances conforming to the rule for properties common to all instances. This process of examining the list of exemplars for common properties and testing by a counterfactual strategy is iterated until the boundaries of the rule are delineated. From a philosophical perspective this strategy employs a Bayesian approach as alternative hypotheses are considered and tested throughout.

1.2.4 The Dual Goal Paradigm

Although first rule-announcement success on the standard 2-4-6 paradigm is poor, Tweney et al. (1980) introduced a manipulation that improved solution rates to over 60%. Tweney et al. asked participants to discover two rules, one producing “DAX” triples (“any ascending number sequence”), the other producing “MED” triples (“any other number sequence”). Evidence for superior performance with these Dual Goal (DG) instructions is remarkably robust (e.g., Farris & Revlin, 1989a, 1989b; Gorman, Stafford, & Gorman, 1987; Tukey, 1986; Wharton, Cheng, & Wickens, 1993), but explanations for the effect remain inconclusive. Tweney et al. suggested that the explanation for the enhanced success rates in their DG paradigm was related to the restructuring of the problem in such a way as to allow participants to use two lines of evidence related to the two hypotheses and that this may be due to one or more of three possible mechanisms:
Chapter 1

Introduction to Hypothesis Testing Behaviour

1. Negative labelling of non-conforming triples in the standard single goal (SG) paradigm interferes with the participant's ability to integrate this information.

2. Distinct labelling of the DAX and MED hypotheses allows them to be processed separately.

3. DG instructions fundamentally alter the logical structure of the task.

Gorman et al. (1987) suggest that the results of their earlier study (Gorman & Gorman, 1984) refute the first possibility as they found that when participants could not ask the experimenter for feedback on their tentative hypotheses, disconfirmatory instructions significantly improved performance. The second alternative is less straightforward to assess in that it is difficult to determine what form this processing might take. Gorman et al. (1987) suggest that instead of separate processing of two rules, DG instructions promote what Wason and Green (1984) call a "unified mental representation". Wason and Green developed a manipulation of the selection task that made negative information more salient, thus facilitating performance. It is possible that DG instructions facilitate a representation of the 2-4-6 task which promotes the salience of disconfirmatory information. As Gorman (1987) notes, participants in the standard form of the 2-4-6 task announce their rule once they have produced a rule sufficient to predict correct triples. If DG instructions are employed, however, this would be the equivalent of stopping when a rule had been produced for the DAX triples but little was known about MED triples.

Gorman (1992a) suggests, therefore, that enhanced performance is achieved on the DG version of the 2-4-6 task by changing the task from one in which participants seek the limits of one rule to one in which they explore the boundaries of two mutually exclusive rules. Some evidence for this view is gained from the protocols of participants in Gorman et al.'s (1987) study. In their experiment participants were required to discover an even more general rule than the usual "ascending numbers" (i.e., "three different numbers"). On post-study questioning one participant stated that he was trying to "change from the first triple in all ways I could to find out what the MED rule was" (Gorman et al., 1987, p. 22), that is, instead of trying to falsify his DAX rule he was searching for a positive instance of the MED rule. In comparing the protocols of participants given DG versus SG instructions who had produced the triple 0-0-0, Gorman notes that in the former case the participants started to explore the boundaries of these types of triple by proposing further examples in which one or more numbers were the same, while the participant in the SG condition concluded this was an exception to the rule and proposed "three consecutive integers, each multiplied by the same number (but not 0)".
Chapter 1

Introduction to Hypothesis Testing Behaviour

Whilst separate processing of the two rules appears to be a plausible explanation of this latter effect, it is important to note, however, that poor performance on the SG task is typically attributed to a failure to seek the limits of the initial rule. Tweney et al. suggest that their third explanation is somewhat akin to Platt’s (1964) heuristic of “strong inference”. The strong inference heuristic is an iterative process which involves the development of alternative hypotheses, the devising of crucial experiments which will exclude one or more of these hypotheses, and the carrying out of these experiments. Platt’s strong inference heuristic thus involves discriminating between two competing hypotheses by the use of critical tests. However, as Gorman et al. (1987) note, the strong inference heuristic is concerned with competing hypotheses, whereas the to-be-discovered rules in the DG paradigm are complementary.

Explaining enhanced performance in the DG paradigm is important, since an account of DG facilitation could lead to substantial progress in understanding poor success rates on the standard paradigm. There are, in fact, several detailed explanations for the DG effect, with each account making different predictions about the factors involved. Prominent accounts are Evans’ (1989) positivity bias theory, Wharton et al.’s (1993) goal complementarity theory, Wharton et al.’s (1995) information quantity theory and Vallée-Tourangeau et al.’s (1995) triple heterogeneity theory. Descriptions and critiques of these theories are presented in the following section.

1.2.5 Theories of DG Facilitation in the 2-4-6 Paradigm

1.2.5.1 Positivity Bias Theory

Evans (1983, 1989) proposed that poor performance on the standard 2-4-6 task can be attributed to the operation of a general positivity bias, a form of selective processing whereby people see positive information as relevant to a task, whereas negative information is dismissed as irrelevant. Evans sees this tendency not as motivational, but rather as a result of preconscious heuristic processes which determine the locus of the participant’s attention. Thus, rather than assuming that participants fail to search for falsifying information it is instead suggested that they fail to attend to it. On the standard Single Goal (SG) 2-4-6 task, people are viewed by Evans as testing positive cases that “match” their overly-restricted hypothesis (e.g., “ascending with equal intervals is right”), but not negative cases (triples such as 2-4-10) that do not match such a hypotheses. As Evans (1983) puts it, “a ‘wrong’ hypothesis is semantically negative and thus overlooked” (p. 144). Interestingly this notion is somewhat akin to Wason’s suggestion (see Tweney, Doherty, Worner, Pliske, Mynatt, Gross et al., 1980) that in the standard paradigm the highly
informative negative triples are called “wrong”, and this may prevent reasoners from using all the available information. At a philosophical level Popper (1989) makes a similar observation when citing the observations of Katz (1937), that a hungry animal divides the world into edible and inedible things, and that a fleeing animal sees escape routes and hiding places, illustrating that objects change according to the needs of the animal. In line with Katz’s observation, Popper suggested that a scientist’s point of view is coloured by their theoretical interests, conjectures and anticipations which form their “horizon of expectations”. Thus in the 2-4-6 task a reasoner will focus on information relating to the target rule and other information will not be attended to.

Evans (1989) argues that when DAX and MED are exchanged for the terms “right” and “wrong” on the DG task, people succeed not because they attempt to disconfirm their DAX hypothesis, but, instead, because they test the MED hypothesis (“not ascending with equal intervals”) with positive triples (e.g., 2-4-10). Because the structure of the DG task means that MED is the complement of DAX, a positive test of MED is effectively a negative test of DAX, which is precisely what is required to eliminate the over-specific DAX hypothesis.

Evans’ account, however, still gives rise to the question of why people in the DG paradigm should concern themselves with testing the MED hypothesis. In addressing this, Evans (1989) clarifies that it is the creation of a positive label (MED) for the negative hypothesis (“not ascending with equal intervals”) that changes people’s conception of the problem, thereby promoting the testing of MED hypotheses. As Evans (1989) explains, “When DAX and MED are substituted for “right” and “wrong”, the two hypotheses appear to have equal standing….Creating a positive label for the negative hypothesis entirely changes subjects’ representation of the task” (p. 52).

Whilst in his updated hypothetical thinking framework Evans (2006) fails to outline specifically how DG instructions lead to facilitated performance on the task, a clear implication is that DG instructions promote the relevance of the MED/“wrong” information, thus the effect occurs in the heuristic processing stage. It could, however, be argued that DG instructions also work at the analytic stage by overcoming the singularity principle. On this interpretation the task moves from a Popperian falsification task towards a Bayesian probability task in which two alternative hypotheses are pitted against each other. Whilst Evans describes the hypothetical thinking framework as a revision of his earlier heuristic-analytic model, this updated model does allow for differing interpretations of the DG facilitatory effect.
1.2.5.2 Goal Complementarity Theory

The goal complementarity theory (Wharton et al., 1993) emphasises three factors that promote DG success: (i) the embedding of the reasoner's initial DAX hypothesis (e.g., "numbers ascending by 2") within the target rule of "any ascending sequence" (Wetherick, 1962); (ii) the complementary nature of the DAX and MED rules, whereby triples fall into two perfectly opposing sets that make up the entire universe of possible triples (i.e., the DAX set reflects the experimenter's rule of "any ascending sequence" whilst the MED set reflects the logical opposite of DAX: "any sequence other than an ascending one"); and (iii) the tendency for people to adopt a "positive test strategy" on the 2-4-6 task (Klayman & Ha, 1987, 1989), which involves generating triples that match the current hypothesis (either DAX or MED) to test its sufficiency. Based on these assumptions, Wharton et al. propose that facilitation arises through two mechanisms: first, whereas in the SG task people only pursue positive tests of the sufficiency of a single rule, in the DG task they alternate in positively testing the sufficiency of both the DAX and the MED rules. Second, a positive test (using a triple like "2-3-10") of the sufficiency of overly broad MED hypothesis (such as "any sequences other than numbers ascending by 2") will lead to unexpected DAX feedback, thereby enabling the participant to realise not only that their current MED hypothesis is too broad, but also that their current DAX hypothesis is too narrow, thus paving the way toward effective rule discovery.

The positive test strategy (Klayman & Ha, 1987, 1989) is identical in its functional consequences to Evans' (1989) notion of positivity bias: both promote the testing of positive exemplars of a single, overly-restricted hypothesis in the SG paradigm, and of the DAX and MED hypotheses in the DG paradigm. Where the positivity bias and goal complementarity theories differ is in their accounts of why people test MED hypotheses at all. Whilst Evans (1989) argues that it is the positive labelling of a negated DAX rule that makes MED hypotheses appear relevant and worth testing (albeit at a preconscious level), the goal complementarity theory holds that people test MED hypotheses because they are directly requested to discover the two rules pertaining to the task.

1.2.5.3 Information Quantity Theory

The information quantity theory of DG facilitation (Wharton et al., 1993) centres around two pieces of evidence: (1) that prior to initial rule announcement SG solvers test more triples than non-solvers (Farris & Revlin, 1989a; Wason, 1960); and (2) that DG instructions evoke the testing of more triples than SG instructions (Gorman et al., 1987). These observations suggest that the quantity of triples produced, irrespective of
their characteristics, may mediate between DG instructions and task success. However, this theory lacks clarity as to the specific mechanisms that promote increased triple testing in DG conditions; it is also questionable why increased testing should be associated with task success. Tweney et al. (1980) and Gorman et al. (Gorman et al., 1987) have gone some way toward addressing these issues, suggesting that switching between the testing of two different rules in the DG paradigm may counter the rigid testing of a single hypothesis, thereby increasing the number of hypotheses examined, the quantity of triples generated, and the likelihood of success.

The information quantity view has some merit, but seems mostly to be a useful description of one aspect of the data surrounding DG superiority (i.e., that DG instructions lead to increased triple production), and is certainly compatible with either the goal complementarity or positivity bias theories. Moreover, these latter accounts have the advantage of providing a clear explanation of why people test more triples in the first place with DG instructions: people do this because they are explicitly requested to discover two rules (the goal complementarity position) or because the positive labelling of negated hypotheses encourages a broader exploration of the hypothesis space (the positivity bias position). It is also compatible with the hypothetical thinking framework (Evans, 2006). Within this framework DG facilitation may be attributed to the overcoming of the singularity principle and reframing of the task in a Bayesian probability model, as two hypotheses are now considered and evidence for them evaluated. Recall that the Bayesian approach is probabilistic in nature and allows for adjustment in the scientist's confidence in their working hypothesis in the light of evidence collected. It seems intuitively appealing that such an approach would lead to more industrious hypothesis testing behaviour until the reasoner is satisfied that confidence in their hypothesis achieves a satisfactory threshold level. Despite the apparent lack of theoretical underpinnings it remains the case that examination of participants' protocols indicates that DG instructions promote both diligent testing of hypotheses and success on the task. Evidence for information quantity theory has largely been through post hoc analysis of participants' protocols with a compelling empirical investigation of the account yet to be published.

1.2.5.4 Triple Heterogeneity Theory

Vallée-Tourangeau, Austin, and Rankin (1995 Exp. 1) explored the link between DG instructions and information quantity in a study where participants had to test 15 triples before rule announcement. This led to 44% solvers with SG instructions and 69% solvers with DG instructions. The fact that DG solvers outstripped SG solvers even though the number of tests was standardised affirms that there is more to the DG effect that the mere quantity of triples tested. Vallée-Tourangeau et al. (1995)
proposed that the crucial factor underpinning DG success is not so much triple quantity as triple heterogeneity (the extent to which triples discriminate between multiple hypotheses), with DG requirements fostering a more flexible and creative exploration of the triple-space. This idea provides a valuable development of the insights concerning multiple hypothesis testing and task success presented by Tweney et al. (1980) and Gorman et al. (1987). It also links with Oaksford and Chater’s (1994) argument that, prima facie, the more hypotheses a person tests the higher will be the probability of success.

Triple heterogeneity theory gains support from Vartanian, Martindale, and Kwiatkowski’s (2003a) study which demonstrated that successful participants on the SG task had higher fluency scores on a measure of creative thinking than unsuccessful participants, and also generated more hypotheses. Intriguingly, however, they noted that whilst the creativity factor was predictive of task success in a stepwise multiple-regression analysis, the number of generated hypotheses did not prove to be a reliable predictor. In fact, the most important factor in rule discovery was the frequency of disconfirming triples generated (i.e., triples that contradicted a participant’s working hypothesis such as stating “6-5-4” to test the hypothesis “numbers ascending by twos”). These findings confirm the value of testing disconfirming triples as highlighted in Wason’s (1960) original analysis of successful 2-4-6 performance, where people who used disconfirmation as their dominant strategy were more likely to announce correct rules on their first guess. However, other data discussed previously suggest that disconfirmation may not always be the most efficient strategy, especially in the initial stages of hypothesis testing (e.g. Tweney, Doherty, Worner, Pliske, Mynatt, Gross et al., 1980).

1.3 Summary

In the preceding sections the literature regarding the philosophy of hypothesis testing has been reviewed alongside that concerned with hypothesis testing behaviour in the laboratory. It is clear that there are several competing views from the philosophy of science as to how science progresses and that evidence from the psychological literature concerning hypothesis testing behaviour is mixed. It is useful at this stage to summarise the status of knowledge concerning hypothesis testing behaviour and also the limits of current research and theorising.

What, then, is the current position regarding research and theory on hypothesis testing behaviour? Evidence from everyday life suggests that in general, humans are very successful reasoners and are able to make successful inductive inferences about
their environment (Poletiek, 2001). However, evidence from empirical studies suggests that, in the laboratory, people are rather less successful at hypothesis testing. Performance in the standard form of the 2-4-6 task (Wason, 1960) has been characterised as being less than diligent, in terms of both the number and types of tests that are produced, with participants using a positive test strategy (Klayman & Ha, 1987; 1989) which leads to announcing of rules that are sufficient but not necessary in scope. However, the introduction of the requirement to find a second rule appears to induce participants to engage in a more diligent and creative exploration of the problem space, with a greater number of tests of a wider variety being performed with a commensurate improvement in performance on the task. Thus it would seem that the key to successful hypothesis testing appears to be to promote this type of diligent hypothesis testing behaviour. The difficulty is that it is unclear why DG instructions promote this type of behaviour, or, indeed, if it is the type and number of tests which are key to success or whether they are merely an artefact of an underlying reconceptualisation of the problem. A second difficulty lies in the dearth of studies which discriminate between competing theories. Although studies have been published which explore the various accounts, close reading of the methodologies shows that many vary on minor points of methodology (e.g., the number of rule announcement opportunities available to participants; minor changes to wording which make the complementarity of rules either explicit or implicit) thus making it difficult to arbitrate between theories. The purpose of this thesis is therefore to carry out a systematic series of studies in which predictions from current theories are tested. Table 1.1 summarises the competing accounts, elucidates the predictions each makes and highlights the crucial test of each account.
### Table 1-1

*Theories of Dual Goal Facilitation, including Predictions and Empirical Tests*

<table>
<thead>
<tr>
<th>Theory</th>
<th>Prediction</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positivity bias theory (Evans, 1989)</td>
<td>Selective processing of information means that negatively valenced feedback will be viewed as irrelevant to the task</td>
<td>Manipulate labelling of feedback (negative vs. positive) across instruction type (SG vs. DG) and compare success rates across conditions</td>
</tr>
</tbody>
</table>
| Goal complementarity theory (Wharton, Cheng, & Wickens, 1993) | Only complementary goals cause DG facilitation, with task success being dependent on the production of at least one positive triple which does not ascend by equal increments | 1. Compare performance using complementary versus non-complementary goal relationships.  
2. Compare the presence of at least one triple which does not ascend by equal increments across conditions and for solvers versus non-solvers |
| Information quantity theory (Wharton et al., 1993) | DG instruction promote the generation of triples, with success being associated with higher number of triples | Control number of triples required across instruction type and compare success rates |
| Triple heterogeneity theory (Vallée-Tourangeau et al., 1995) | DG complementarity is not critical for DG facilitation, which instead depends upon the production of a wider variety of triples types | Analyse the variety of triple types across conditions and for solvers versus non-solvers |
Chapter 2
General Methodology and Analysis

2.1 Introduction

As noted above, minor variations in methodologies in the 2-4-6 paradigm may well impact in important ways on the behaviour of interest. In order to avoid possible confounds in the present thesis it was decided to hold constant some key aspects of the experimental methodology throughout all reported experiments. To prevent unnecessary repetition these standardised aspects of the present research are outlined below, along with a rationale for their employment. For the same reason details of some of the analytic methods that were deployed in examining the empirical data in this thesis are also provided below.

2.2 Scoring Success on the DG Task

In scoring participants as being successful or unsuccessful on the DG version of the task it was only necessary for their initial rule announcement to relate to the OAX rule. Any announcements for the MED rule were disregarded, as is standard practice in research with DG task variants (e.g. Vallée-Tourangeau et al., 1995; Wharton et al., 1993).

2.3 Coding of Triples for Analyses of Hypothesis Testing Strategies

Throughout this programme of study it was necessary to analyse participants’ hypothesis testing strategies using both quantitative and qualitative analyses of the triples generated. It is straightforward to obtain measures of total triple quantity (i.e., the cumulative count of a participant’s generated triples), and of total triples receiving negative feedback (i.e., the cumulative count of a participant’s generated triples receiving “MED” or “does not fit” responses from the experimenter). Such measures are computed from the datasets on a by-condition basis.

Obtaining a measure of the variety of triples generated by participants is more complex, and involves recourse to techniques pioneered by Vallée-Tourangeau et al. (1995), which focus on two main classes of triple referred to as variable positives (posvars) and negative types (negtypes). Posvars are triples such as 2-8-20 that receive positive feedback (DAX or “fits”) but that do not increase by a constant number as in
the 2-4-6 exemplar (described as having a "constant positive" form). Thus, if the numbers that make up a triple are denoted by the letters a, b, and c, then a posvar is any triple in which \((b - a) \neq (c - b)\), whereas a constant positive is any triple where \((b - a) = (c - b)\). Posvar scores are computed for each participant, and are simply a cumulative count of the number of posvar triples produced.

Negtypes reflects a measure of the heterogeneity of negative triples. There are eight possible types of triple that could receive negative feedback, such as descending triples and identical-number triples. The possible set of negtypes is captured by the following rules: (1) \(a > b > c\); (2) \(a = b = c\); (3) \(a > b < c\); (4) \(a < b > c\); (5) \(a = b < c\); (6) \(a < b > c\); (7) \(a > b = c\); (8) \(a < b = c\). To obtain a negtype score for each participant, the number of distinct types of negative triple that they produced is counted. Thus, if a participant generates five negative triples of the same kind (say the decreasing triples of the \(a > b > c\) form), then their negtype score equals one. If the participant generates three negatives, two decreasing and one \(a < b > c\) "hill" kind, then their negtype score would be two. These indices of triple heterogeneity are used throughout the present research to provide insight into participants' hypothesis testing strategies.

2.4 Restricting Opportunities for Rule Announcement

One way in which the methodology employed in the 2-4-6 task varies is in the number of rule announcements which are permitted. In his original study Wason (1960) asked participants to test triples until they were confident of the rule, at which point they should announce their hypothesis. If this hypothesis was incorrect, they were informed of this and asked to continue generating and testing triples; thus participants were encouraged to continue with the task until either they had discovered the rule or they had reached the time limit.

Many researchers followed this methodology and allowed multiple rule announcements (e.g., Rossi, Caverni & Girotto, 2001; Tukey, 1986; Vallée-Tourangeau & Payton, 2008). However, by allowing multiple rule announcements participants now have a second (and more powerful) method of testing their hypotheses, that of asking the researcher for feedback on their hypothesis (Gorman, 1992). This additional method lacks ecological validity in terms of modelling scientific discovery in that scientists are not able to appeal to an outside authority (i.e., the experimenter) to check the truth of a working hypothesis, and as such undermines the task as a method of investigating scientific enquiry.

A second issue with allowing multiple rule announcements is that in this thesis participants' hypothesis testing strategies were delineated by the coding of triples as
outlined above. Allowing a secondary method of hypothesis testing may impact on triple generation patterns and therefore undermine the analysis of triple generation profiles as a method of uncovering hypothesis testing strategies. For these reasons, and to maintain a level of comparability across the experiments to be reported in this thesis, it was decided to employ the methodology of Gorman (1992b) in which participants are allowed a single opportunity to announce the target rule.

2.5 Reporting of Effect Sizes

In line with APA (2001) recommendations, effect sizes are reported wherever possible. As effect sizes are independent of sample size such a strategy enables comparison of effects across studies and analyses thereby providing a measure of the degree to which the phenomenon of interest is present in the population. There are a variety of possible effect size statistics available to the researcher, but effect size statistics reported here follow the recommendations of Cohen (1977; 1988), that is, for t tests and Mann Whitney tests Cohen’s $d$ is reported, for chi square analyses, $w$, and for two-way ANOVAs, $\eta^2$. Throughout the thesis the convention of regarding effect sizes as small, medium or large as described by Cohen is also adhered to, although note is taken of Cohen’s caution that such descriptors may inherently be misleading given that qualitative concepts such as large are sometimes understood as absolute and other times as relative. Table 2.1 gives Cohen’s descriptions of small, medium and large effect sizes for effect size statistics employed in this thesis.

Table 2-1

<table>
<thead>
<tr>
<th>Test</th>
<th>Index</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Independent t test</td>
<td>$d$</td>
<td>.20</td>
<td>.50</td>
<td>.80</td>
</tr>
<tr>
<td>Two-way ANOVA</td>
<td>$\eta^2$</td>
<td>.01</td>
<td>.059</td>
<td>.138</td>
</tr>
<tr>
<td>Chi square test of independence</td>
<td>$w$</td>
<td>.10</td>
<td>.30</td>
<td>.50</td>
</tr>
<tr>
<td>Mann Whitney U</td>
<td>$d$</td>
<td>.20</td>
<td>.50</td>
<td>.80</td>
</tr>
</tbody>
</table>

Notes: Conversion from $f$ to $\eta^2$ by Clark-Carter (2004), and $d$ in Mann Whitney U converted from $z$ test using calculation recommended in Clark-Carter (2004).
2.6 Path Analysis and Logistic Regression

Some of the theories to be explored in this thesis predict the effect of one variable on the outcome variable by a third, mediating variable. For this reason it has been necessary to explore such mediation effects using path analysis (Baron & Kenny, 1986). However the outcome variable in these analyses is dichotomous (solver/non-solver; presence/absence of triple type) which required analysis using logistic regression. However, because the quantitative logic of path analysis does not work effectively with logistic regression, it was necessary to use linear regression for this mediation analysis. This is not ideal given the dichotomous nature of the relevant variables, but in such cases comparisons of obtained p values for logistic and linear regressions are presented to illustrate the high degree of similarity in the statistical outcomes of these two approaches, and thereby to validate the use of the linear-regression procedure with the dataset.
3.1 Introduction

The aim of this first experiment was to develop a crucial test to arbitrate between the Evans (1989) positivity bias and Wharton et al.’s (1993) goal complementarity views of the 2-4-6 task as measured in terms of successful initial rule announcement. Recall that the positivity bias account attributes facilitation to the overcoming of preconscious heuristic processes which cause the reasoner to attend to positive rather than negative information. Relabelling the negatively valenced “wrong” feedback to the positively valenced “MED” feedback is proposed to cause the reasoner to overcome this tendency and promotes the testing of the MED hypothesis leading to successful delineation of the DAX and MED boundaries respectively. The goal complementarity account, however, suggests that facilitation is caused by a combination of the embedding of the initial hypothesis in the target hypothesis, the mutually exclusive but exhaustive set boundaries of the two rules, and the tendency to positive testing (Klayman & Ha, 1987, 1989). The combination of these three factors, along with the requirement to discover two rules, leads the reasoners to alternatively test the DAX and MED hypotheses leading to successful rule discovery.

Despite the conceptual overlap between these accounts a crucial test is possible (cf. Wharton et al., 1993). The key difference between these theories is that whilst for Evans (1989) DG facilitation is tied to the use of a positive label (MED) to denote a negative hypothesis (“not ascending with equal intervals is wrong”), for Wharton et al. (1993), what is critical is the explicit instruction for participants to discover the MED as well as the DAX rule in the DG paradigm. One consequence of the positivity bias theory is that it would predict that participants given positively labelled hypotheses (DAX and MED) should perform better than participants given mixed valence hypotheses of the form “fits” (positively labelled) and “does not fit” (negatively labelled) regardless of whether SG or DG instructions are employed. In contrast, the goal complementarity theory would predict that participants given the task of discovering two complementary hypotheses should be more successful than those seeking a single hypothesis, regardless of the way in which these rules are labelled.
Chapter 3

Experiment I: Comparing Positivity Bias and Goal Complementarity Theories

In addition to arbitrating between the goal complementarity theory and the positivity bias account of DG facilitation a secondary aim of this study was to explore the triple heterogeneity (Vallee-Tourangeau et al., 1995) and information quantity (Wharton et al., 1993) accounts of DG facilitation by determining whether the number and characteristics of triples produced mediate successful rule discovery. The triple heterogeneity account proposes that facilitation is a result of the wide exploration of the problem space as indexed by the variety of triples generated which is instrumental to success on the task, whereas information quantity theory proposes that DG instructions promote increased triple production which in turn leads to successful performance.

Thus for Experiment I it was decided to independently manipulate the mode of instruction and the linguistic labelling of feedback to arbitrate between the positivity bias and goal complementarity accounts of the DG facilitation effect and to undertake both quantitative and qualitative analysis of the triples participants produced in order to help discriminate between the competing accounts of DG performance. Predictions derived from the positivity bias account suggest that in terms of task performance participants given DAX versus MED feedback will be more successful than those given “fits” versus “does not fit” feedback, while the goal complementarity account suggests that performance will be better in the DG variant regardless of the linguistic labelling of the feedback. In terms of triple generation profiles the predictions for the different theories can be summarised as follows: (1) the goal complementarity theory predicts that production of a posvar triple will be associated with success; (2) the information quantity account suggests that DG instructions will promote generation of a greater number of triples and that this will be associated with success on the task; and (3) the triple heterogeneity theory emphasises that success on the DG variant is associated with a wider range of triples being generated and therefore predicts a higher negtype score for both those in the DG condition and successful solvers.

3.2 Method

3.2.1 Participants

Sixty undergraduates from Derby University participated in the experiment for course credit. They had not received any prior teaching on the psychology of reasoning.
3.2.2 Design

The experiment involved random allocation of 15 participants to each of four conditions. Two conditions presented DG instructions and included a linguistic labelling manipulation whereby participants were either asked to discover DAX and MED rules (positive label condition) or "fits" and "does not fit" rules (mixed label condition). The remaining two conditions involved SG instructions and simply acted as controls for the presented feedback in the DG conditions (i.e., feedback was either of the form DAX-MED or "fits"-"does not fit").

3.2.3 Procedure

Participants were tested in groups of up to four in a quiet laboratory. The following safeguards avoided social contamination of individual results: (1) participants sat well apart in screened cubicles, (2) participants wrote down triples and rules and were provided with written feedback; and (3) no spoken communication was permitted during the experiment (see Gorman & Gorman, 1984; Gorman et al., 1987 for effective use of group testing with the 2-4-6 task). The SG instructions referred to a unique rule that needed to be discovered, and stated, "I have in mind a rule that specifies how to make up sequences of three numbers (triples), and your task is to discover this rule". In what is subsequently referred to as the SG-Fits condition, participants were asked to discover the target rule by generating triples that they would then be told either fitted or did not fit the rule that the experimenter had in mind. In the SG-DAX condition, participants were told that triples that fitted the rule were called DAX triples and those that did not fit the rule were called MED triples. It was explained to participants that on generating a triple they would be informed as to whether it was a DAX or a MED type.

The DG instructions emphasised that there were two rules to be discovered, "Your task is to discover this rule, and also a second rule for categorising the triples that do not fit my rule". In the DG-Fits condition participants were asked to generate triples that would be classified in terms of whether they fitted or did not fit the rule. In the standard DG task (DG-DAX) participants were informed that triples that fitted the rule were called DAX and those that did not fit the rule were called MED. They were instructed to produce further triples that the experimenter would classify as being DAX or MED.

Participants in all conditions were given "2-4-6" as the example triple. All participants were also provided with an answer sheet and were asked to write "2-4-6" on the first row and either "fits" or "DAX" in the feedback column, as appropriate. They were instructed that they could produce as many triples as they wished, and that
when they were sure of the rule(s) they should write it/them on the answer sheet. In line with Gorman et al., (1992a) participants were allowed only one guess at the rule or rules.

3.3 Results

3.3.1 Solution success across conditions

Table 3.1 shows the frequency of correct and incorrect initial rule announcements in each of the experimental conditions. A number of findings are apparent. First, nearly four times the number of participants in the DG conditions announced the correct rule when compared with the SG baseline conditions. Second, the linguistic labelling manipulation associated with the DG rules had little impact on rule discovery, with nearly equal numbers of participants finding the sought for rule in the positive label and mixed label conditions. Third, the distribution of initial solvers and non solvers was identical across the SG conditions, indicating that the labelling of feedback (DAX-MED vs. “fits”-“does not fit”) had no effect on the likelihood of successful rule announcement.

A contingency table chi square analysis was performed on the frequencies of correct and incorrect announcements, collapsing across the two SG conditions and the two DG conditions (see Table 3.1). This revealed a highly reliable effect of Goal Requirement (SG vs. DG), \( \chi^2 (1, N = 60) = 19.29, p < .001, \omega = .57 \). In terms of correct rule announcements the results arbitrate in favour of the predictions of the goal complementarity account of DG facilitation (e.g. Wharton et al., 1993) and against the predictions of the positivity bias account (e.g. Evans 1989).

3.3.2 Quantity and variety of triples generated across conditions

Table 3.1 presents mean scores by condition for the number of posvars and negtypes generated in the present study, as well as means for the total quantity of triples generated and the number of triples receiving negative feedback. To examine whether the experimental manipulations had an effect on any of these indices, a series of two-way analysis of variance (ANOVA) tests were undertaken where the factors were Goal Requirement (SG vs. DG) and Linguistic Labelling of Feedback (i.e., the experimental controls involving the linguistic balancing of feedback across the SG and DG tasks allowed for a comparison between conditions where DAX-MED feedback had been given vs. conditions where “fits”-“does not fit” feedback had been presented).
Table 3-1

*Frequency of Correct Initial Rule Announcements (Solvers vs. Non solvers), Mean Numbers of Total Triples Produced, Types of Triples Produced (Posvars or Negative Types), and Triples Receiving Negative Feedback, by Condition*

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Solvers</th>
<th>Non-Solvers</th>
<th>Total Triples</th>
<th>Posvars</th>
<th>Negative Types</th>
<th>Negative Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG—DAX</td>
<td>15</td>
<td>3</td>
<td>12</td>
<td>7.60 (1.62)</td>
<td>0.33 (0.16)</td>
<td>0.33 (0.16)</td>
<td>0.73 (0.33)</td>
</tr>
<tr>
<td>SG—Fits</td>
<td>15</td>
<td>3</td>
<td>12</td>
<td>5.87 (0.71)</td>
<td>0.53 (0.35)</td>
<td>0.80 (0.37)</td>
<td>0.93 (0.41)</td>
</tr>
<tr>
<td>DG—DAX</td>
<td>15</td>
<td>12</td>
<td>3</td>
<td>10.27 (1.63)</td>
<td>1.13 (0.27)</td>
<td>1.20 (0.22)</td>
<td>2.67 (0.50)</td>
</tr>
<tr>
<td>DG—Fits</td>
<td>15</td>
<td>11</td>
<td>4</td>
<td>8.33 (0.83)</td>
<td>1.07 (0.28)</td>
<td>0.93 (0.21)</td>
<td>1.40 (0.32)</td>
</tr>
</tbody>
</table>

*Note:* SG = Single Goal; DG = Dual Goal. Standard error of the mean in parentheses.
Consistent with the task success analyses already reported, Linguistic Labelling of Feedback produced no reliable differences on any of the measures of triple quantity or type. With regard to Goal Requirement, however, there were significant main effects on total number of triples produced, $F(1, 56) = 4.09, p < .05, \eta^2 = .066$, number of triples receiving negative feedback, $F(1, 56) = 9.10, p < .01, \eta^2 = .129$, and number of posvars, $F(1, 56) = 5.86, p < .05, \eta^2 = .094$, thus the effect sizes reported here are in the medium to large range. The difference in the number of negative types produced across SG and DG conditions also approached significance, $F(1, 56) = 3.96, p = .052, \eta^2 = .063$. There were no significant interactions between factors for any of the measures (all $F$s < 1). These results underline the importance of DG instructions as a determining factor in engendering quantitative and qualitative changes in triple generation on the 2-4-6 task. As such, the findings again favour the goal complementarity view of DG facilitation.

### 3.3.3 Presence of triple types and solution success

Although the previous analyses indicate an effect of DG instructions on measures of triple quantity and type, they do not reveal the importance of generating specific types of triple for actual task success. Further analyses in which the production of either at least one posvar or at least one negative type was crossed with success were therefore pursued. Such analyses are important for a simple but critical reason. The point is, both the goal complementarity and the positivity bias accounts place a central emphasis on the role of posvars in engendering successful rule announcement, that is, both accounts claim that once people generate a posvar (e.g., 2-4-10) that receives DAX or "fits" feedback they should be able to ascertain immediately that the DAX or "fits" rule is broader than the initially hypothesised form. On the other hand, both theories are silent as to the value of negative testing in facilitating task success.

To examine the association between generation of posvars and success a contingency table was produced in which the production by a participant of at least one posvar or at least one negative type was crossed with success were therefore pursued. Table 3.2 reveals that such an association is indeed evident, with substantially more participants who produced a posvar making a correct rule announcement compared to those who did not produce a posvar. A chi squared analysis confirmed the reliability of this observation, $\chi^2 (1, N = 60) = 13.14, p < .001, w = .468$. 

48
Table 3-2

*Frequency of Correct Rule Announcements by Presence versus Absence of Posvar Triples, and by Presence versus Absence of Negative Triples*

<table>
<thead>
<tr>
<th>Presence of triple type</th>
<th>N</th>
<th>Solvers</th>
<th>Non-Solvers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posvar present</td>
<td>25</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>Posvar absent</td>
<td>35</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>29</td>
<td>31</td>
</tr>
<tr>
<td>Negative triple present</td>
<td>34</td>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>Negative triple absent</td>
<td>26</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>29</td>
<td>31</td>
</tr>
</tbody>
</table>

Table 3.2 also shows a contingency table in which production of at least one negative triple was crossed with success. Here the association is even more striking than in the case of posvar production, with there being only one instance of a participant correctly announcing the rule but not producing a negative triple. In contrast, of the 34 participants who did produce a negative triple, 28 solved the task. A chi square analysis indicated these differences were highly significant, $\chi^2 (1, \ N = 60) = 36.36, p < .001 \ \text{w} = .778$. These latter findings - that production of at least a single triple receiving negative feedback is more closely associated with success than production of a single posvar - seem paradoxical. Since most incorrect rule announcements are of the “numbers ascending by equal intervals” type it would appear that only production of posvars could lead to the falsification of overly-restricted hypotheses, whereas production of negative triples would seem of little obvious value for rule discovery.

To clarify the association between the triple type variables and task success the dataset was modelled using logistic regression. An initial model using negative triple (present vs. absent) as the predictor variable, and success (solver vs. non solver) as the outcome variable, revealed that negative triple was a highly reliable predictor of task success, $B = 4.76, \ \text{Wald} = 18.23, p < .001$. A second model using posvar (present vs. absent) as the predictor, and success (solver vs. non solver) as the outcome variable again revealed this predictor to be reliable, $B = 2.07, \ \text{Wald} = 11.91, p < .001$. A final
model was assessed in which negative triple and posvar were regressed onto success in a hierarchical manner to examine whether the posvar predictor had an effect additional to the negative triple predictor. Negative triple was selected as the first predictor due to the considerably higher overall chi square value obtained for Model 1, $\chi^2 = 42.95, p < .001$, than for Model 2, $\chi^2 = 13.68, p < .001$. This third model showed that negative triple continued to be a highly reliable predictor of task success, $B = 4.58, \text{Wald} = 15.71, p < .001$, but with negative triple controlled for the posvar predictor failed to achieve significance, $B = 1.72, \text{Wald} = 3.55, p > .05$. This finding suggests that the production of posvar triples may have a limited association with successful rule discovery on the 2-4-6 task in comparison to the more striking influence that the generation of negative triples seems to have.

Since production of negative triples appeared to have such a strong association with task success it was important to examine what specific property of negative triples might underpin this phenomenon. It was clear upon scrutinising the negative triples which participants generated that the majority were “descending” in nature, and that most other types of negative triples were produced infrequently (only 10 participants produced a negative type distinct from the descending $a > b > c$ form). For this reason all negative triples apart from those of the $a > b > c$ type were collapsed into a single pool. Separate analyses were then pursued comparing: (1) the effect of producing versus not producing at least one descending triple on task success, and (2) the effect of producing versus not producing at least one other type of negative triple on task success. Logistic regression revealed that participants generating a descending triple were 142 times more likely to solve the task than those not producing a descending triple, $B = 4.98, \text{Wald} = 19.47, p < .001$. Production versus non production of at least one other type of negative triple was also predictive of success, $B = 2.60, \text{Wald} = 5.67, p = .017$, although clearly not to such a marked degree as production versus non production of descending triples.
Table 3-3

**Summary of Mediation Analysis of the Relationship Between Goal, Production of at Least One Descending Triple, and Success, Using Linear Regression (Including a Comparison of the p Values Obtained Using Linear and Logistic Regression)**

<table>
<thead>
<tr>
<th>Variables in Regression Analysis</th>
<th>B</th>
<th>Beta</th>
<th>SE(B)</th>
<th>Linear Regression</th>
<th>Logistic regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal onto Solver</td>
<td>.567</td>
<td>.567</td>
<td>.108</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Goal onto Descending Triple</td>
<td>.500</td>
<td>.503</td>
<td>.113</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Goal and Descending Triple onto Solver</td>
<td>Goal</td>
<td>.215</td>
<td>.215</td>
<td>.086</td>
<td>.015</td>
</tr>
<tr>
<td></td>
<td>Descending</td>
<td>.703</td>
<td>.700</td>
<td>.086</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>


3.3.4 Path Analysis of Mediator Effect

The results of the logistic regression analyses are instructive, but the role of production of a descending triple in mediating between goal requirement (SG vs. DG) and task success could better be illustrated using path analysis (Baron & Kenny, 1986). However, as noted in Chapter 2, because the quantitative logic of path analysis does not work effectively with logistic regression, it was necessary to use linear regression for this mediation analysis. This is not ideal given the dichotomous nature of the relevant variables, but comparisons of obtained p values for logistic and linear regressions (Table 3.3) are presented to illustrate the high degree of similarity in the statistical outcomes of these two approaches, and thereby to validate the use of the linear regression procedure with the present dataset. Table 3.3 also summarises the results of this path analysis of the relationship between Goal Requirement (SG vs. DG), Descending Triple (present vs. absent), and Success (solver vs. non-solver) in the 2-4-6 task. The amount of mediation relating to the production of a descending triple was large (0.352). A Sobel test revealed that this mediating effect was highly reliable, Goodman (I) = 3.87, p < .001.

3.4 General Discussion

Task success findings reported here indicate that the DG superiority effect in the 2-4-6 task cannot be attributed to labelling a negatively valenced “does not fit” hypothesis as a positively valenced “MED” hypothesis: Participants instructed to discover two rules performed significantly better than those in SG conditions, regardless of whether the two rules were defined as DAX-MED or as “fits”-“does not fit”. Such findings support key elements of Wharton et al.’s (1993) goal complementarity account of the facilitatory effect of DG instructions, and run counter to Evans’ (1983; 1989) positivity bias account which proposes that people selectively attend to positively labelled information at the expense of attending to potentially useful information that is negatively labelled. Although the idea of positivity bias affecting hypothesis testing is appealing (cf. Ball, Lucas, Miles, & Gale, 2003), it seems that the concept of a generalised positivity bias cannot easily extend to an explanation of behaviour on Wason’s 2-4-6 task.

Despite the solution success evidence for goal complementarity theory it could be argued that a strong version of this theory is undermined by evidence that strict rule complementarity is unnecessary for rule discovery. For example, Vallée-Tourangeau et al. (1995) ran DG conditions that explicitly suggested a non-complementary representation of the DAX and MED rules. In one condition participants were told
that triples could be DAX, MED or neither, and in another they were told that triples could be DAX, MED, or both. With these manipulations 80% of people still discovered the DAX rule on initial announcement. One weakness of Vallée-Tourangeau et al.’s (1995) study was that although the apparent relationship between the DAX and MED rules was manipulated such that they were not represented as complementary, the reality was that (unbeknownst to participants) the two rules actually remained logically complementary and feedback other than DAX or MED was never given.

As for information quantity theory (Wharton et al., 1993) these results support the view that DG instructions promote increased test generation when compared with SG instructions. Likewise, in relation to triple heterogeneity theory (Vallée-Tourangeau et al., 1995), increased generation of triples receiving negative feedback and increased posvar triples in DG conditions relative to SG ones was observed. It thus seems that DG facilitation is mediated by both quantitative and qualitative changes in triple testing behaviour. As noted previously, however, information quantity and triple heterogeneity theories are perhaps more descriptive than explanatory in emphasis, and may be better subsumed within other contemporary accounts of the 2-4-6 task.

The analyses of triple types across condition and success also offer some support for the hypothetical thinking framework (Evans, 2006). Recall that it was observed that this framework attributes poor performance on the SG variant of the task to a combination of the relevance principle at the heuristic processing level leading to the production of an overly-restricted DAX rule, and the satisficing principle at the analytic processing level which leads to acceptance of this overly-restricted hypothesis as being “good enough”. It was suggested, however, that the third principle of the hypothetical thinking framework, singularity, may impact on DG facilitation; that the instruction to find the second rule may in fact lead to overcoming the tendency to consider only one hypothesis at a time. It was argued that such an explanation would lead to a more Bayesian approach to hypothesis testing as the two alternative hypotheses are considered. It would be expected that this would lead to more extensive testing of the two hypotheses as the reasoner’s confidence varied according to the strength of the evidence collected until a criterion confidence level is achieved. Thus predictions derived from this approach would be both increased triple production and triple variety in the DG variant of the task as participants are required to test two hypotheses. These predictions are supported by the findings of this study.

The final set of analyses also reveal a hitherto unremarked phenomenon. It has long been noted that people who solve the 2-4-6 task tend to produce more triples as
well as a greater proportion of negative triples (Wason, 1960). It has also been demonstrated more recently that solvers generate a greater variety of triples (e.g. Vallée-Tourangeau et al., 1995). What has not previously been shown, however, is that it is the production of at least a single negative triple that is so closely associated with success on the task. Indeed it remains possible that other indices of success such as the total number of triples produced or overall triple variety may well be mediating factors through which the critical negative triple is produced as a result of task manipulations. This is an area which would seem to require closer investigation.

The basic observation that negative triple production is so closely related to task success, does, at first sight, appear rather paradoxical. The point is, that given the typically overly restrictive hypotheses which participants form, it seems intuitively obvious that it should be the production of the discriminatory posvars (rather than negative triples) that would be most strongly associated with task success. Although these results do indicate that posvar generation is significantly linked to correct initial rule announcements on the task, it remains striking that the production of negative triples is even more predictive of task success. Why might this be the case?

One possibility is that the production of a descending triple (and its associated MED or “does not fit” feedback) somehow makes the general dimension of ascending numbers appear to be relevant to the target DAX or “fits” rule. The concept “descending” may have this effect by facilitating the establishment of a salient contrast class that promotes an insight into the potential scope of the target rule. Closer investigation of the precise role of negative triples in facilitating task success - perhaps through the invocation of clear contrast sets within the space of possible triples - would, therefore, appear to be essential. To achieve this a finer-grained system of codifying the triples that participants produce may be required.

Overall, it could be argued that this study has progressed an understanding of DG facilitation effects in the 2-4-6 task. The basic task success measures provide little support for a positivity bias view of the DG effect (e.g. Evans 1989), as increased levels of task success can arise even with negatively labelled rules, so long as participants are still instructed to discover two rules. Although this latter result provides support for the goal complementarity view that direct requests to discover two rules are central to DG facilitation (Wharton et al., 1993), it has also been argued that there is other evidence that calls into question the necessity of having logically complementary rules within DG manipulations. Indeed, the idea that DG instructions may promote production of descending triples seems to be more pivotal to enhanced performance of DG variants of the 2-4-6 task, an observation which runs contrary to
Experiment 1: Comparing Positivity Bias and Goal Complementarity Theories

the predictions of goal complementarity theory, and as such is a notion which needs to be explored more fully.
Chapter 4
Experiment 2: Investigating Information Quantity Theory

4.1 Introduction

While Experiment 1 demonstrated that the positivity bias account (Evans, 1983; 1989) is not a convincing explanation of DG facilitation, the analyses did little to discriminate between other competing accounts of the effect. It was noted that as predicted by the information quantity account (Wharton et al., 1993) solvers did indeed produce more triples than non-solvers. In addition, as triple heterogeneity account predicts (Vallée-Tourangeau et al., 1995), the triples generated by solvers were seen to be of a greater variety than those of non-solvers. It was also noted that some support for the goal complementarity account (Vallée-Tourangeau et al., 1995) was obtained in that the request to discover two complementary rules appeared to promote success on the task. Evidence for this was, however, somewhat equivocal in that, in contrast to the predictions of goal complementarity, the production of a descending triple was more closely associated with success on the task than was the production of a posvar triple. However, this is the first time that such an association has been reported and, as such, this finding needs to be replicated.

These observations led to the second experiment in this series of studies, which instituted a manipulation that aimed to test directly the information quantity account of DG facilitation. Although Wharton et al. (1993, Exp. 3) tested the information quantity account of the task against the goal complementarity account, their manipulation required participants to generate at least 5 triples prior to rule announcement. However, participants were actually allowed to generate more than 5 triples, and, as such, triple generation was therefore not strictly controlled. The results of the study showed that individuals in the DG condition did, in fact, produce more triples than those in the SG condition, and that they were also more successful. However, while it is the case that DG instructions both increase triple production and improve performance, these observations both rely on post hoc analyses of performance, thus a claim of a causal mechanism cannot be made. Indeed it may be that some other mechanism underlies both of these observations.

A stronger test of the information quantity hypothesis would be to manipulate the number of triples participants are required to produce across instruction type and compare success rates. If merely quantitative changes in triple production are
sufficient to promote success on the DG variant of the task then simply requiring participants to produce more triples would have a favourable effect on performance regardless of instruction type. Wharton et al. do not provide analyses of the types of triples produced and it is thus unclear whether there were qualitative differences in the types of triples produced by their DG and SG participants, which would indicate differing hypothesis testing strategies. Vallée-Tourangeau et al. (1995, Exp. 1) implemented a manipulation which addressed these issues when they required participants to generate exactly 15 triples in both their SG and DG conditions before announcing the rule. Analysis of triples produced in their study showed that participants in the DG condition were more successful than those in the SG goal condition, that on average they produced more than twice the number of negative triples, more posvar triples and more types of negative triple. These results suggest that DG instructions do indeed cause qualitative changes in triple production and it might therefore be that increased triple production is merely an artefact of this.

One criticism of Vallée-Tourangeau et al.'s study is that there are few reported instances of participants producing 15 triples with either SG or DG instructions, and it is a possibility that the requirement for such a large number of triples affected the hypothesis testing strategies of their participants. For this reason it was decided to replicate the study with the additional requirement that participants produce either 6, 12 or 18 triples. These numbers of triples were selected after an informal review of the rule discovery task literature which showed that in the majority of studies the mean number of triples produced before the first rule announcement was just under six when using SG instructions (e.g. Wason, 1960; Wharton et al., 1993). It is more difficult to ascertain the typical number of triples produced in the DG condition, but Experiment 1 indicated that when responding to standard DG instructions the mean number of triples generated was just over 10. With 11 being an unusual request, and to avoid alerting participants to the relevance of the number of triples demanded, it was decided to ask for 12 triples as being representative of DG triple production. The 18 triple condition was introduced as a final comparator on the basis that if the information quantity has a role to play in dual goal facilitation then participants in this group would be more successful than those asked to produce six or twelve triples. As with Experiment 1, a series of post hoc analyses was planned to arbitrate between other competing theories of DG facilitation, including the notion that contrast class cues may have a role to play, as mentioned in the previous chapter. Table 4.1 summarises the various theories, the predictions that can be derived from them and the planned analyses.
Table 4-1

*Theories of Dual Goal Facilitation Examined in the Present Study, Including Predictions and Empirical Tests*

<table>
<thead>
<tr>
<th>Account</th>
<th>Features</th>
<th>To Test</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information quantity theory (Wharton et al., 1993)</td>
<td>DG instructions promote triple generation.</td>
<td>Manipulate number of triples across instruction type.</td>
<td>Success rates will rise with requirement to produce more triples regardless of instruction type</td>
</tr>
<tr>
<td></td>
<td>Task success is dependent on producing a high number of triples regardless of condition.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal complementarity theory (Wharton et al., 1993)</td>
<td>Only complementary goals cause DG facilitation.</td>
<td>Compare presence of posvars across conditions as well as for solvers versus non-solvers</td>
<td>Production of a posvar triple mediates between type of instruction and success on the task</td>
</tr>
<tr>
<td></td>
<td>Task success dependent on the production of at least one posvar triple</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple heterogeneity theory (Vallée-Tourangeau et al., 1995)</td>
<td>DG facilitation depends upon the production of a wider variety of triples types</td>
<td>Analyse the variety of triple types across conditions as well as for solvers versus non-solvers</td>
<td>Variety of triples produced mediates between type of instruction and success on the task</td>
</tr>
<tr>
<td>Contrast class theory (Experiment 1)</td>
<td>DG instructions promote production of descending triple</td>
<td>Compare presence of descending triple across instruction mode and for solvers versus non-solvers</td>
<td>Production of a Descending triple mediates between instruction mode and success on the task</td>
</tr>
<tr>
<td></td>
<td>Descending triple generation associated with task success</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2 Method

4.2.1 Participants

Ninety sixth form students from local schools and colleges (aged 18 or over) took part voluntarily in the study. No participants had received training in logic or reasoning before the experiment.

4.2.2 Design

A fully between participants design was employed that manipulated two factors: Goal (Single Goal vs. Dual Goal), and Number of Triples (6, 12 or 18). Fifteen participants were randomly assigned to each of the six resulting conditions.

4.2.3 Materials and Procedure

Participants were tested in groups of two to four, with each participant producing written triples and receiving written feedback. Participants were provided with a sheet on which to record triples, feedback and their best guess at the rule after producing the required number of triples.

Single Goal (SG) instructions referred to a unique rule and are reproduced here in full: "I have in mind a rule that specifies how to make up sequences of three numbers (triples), and your task is to discover this rule. Triples that fit my rule are called DAX triples, and those that do not fit my rule are called MED triples. To start you off, I can tell you that 2-4-6 is a DAX triple. In order to discover the DAX rule you should produce further number triples, and I will tell you whether they are DAX triples, or whether they are MED triples". They were further instructed to test 6, 12 or 18 triples after which they would be asked to write what they thought was the rule for producing DAX triples.

Dual Goal (DG) instructions differed from those for the SG condition by the insertion of the phrase: "...your task is to discover this rule, and also a second rule for categorising the triples that do not fit my rule". DG participants were instructed to test 6, 12 or 18 triples according to condition, after which they would be asked to write what they thought was the rule for producing DAX and MED triples.
4.3 Results and Discussion

4.3.1 Task Success across Conditions

Table 4.2 shows the percentage of correct and incorrect initial rule announcements in each of the experimental conditions. A number of findings are apparent. First, contrary to the prediction of the information quantity account (Wharton et al., 1993) nearly twice the number of participants in the DG condition solved than in the SG condition. Second, success rates did not improve as the number of triples generated increased. Indeed, there is some indication that in this study 12 triples was the optimum number of triples to be produced.

Table 4-2

<table>
<thead>
<tr>
<th>Instruction Type</th>
<th>6</th>
<th>12</th>
<th>18</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Goal</td>
<td>20</td>
<td>53</td>
<td>26</td>
<td>33</td>
</tr>
<tr>
<td>Dual Goal</td>
<td>46</td>
<td>73</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>Total</td>
<td>33</td>
<td>63</td>
<td>43</td>
<td></td>
</tr>
</tbody>
</table>

A chi-square test of independence was performed on the frequencies of correct and incorrect announcements, collapsing across the three SG conditions and three DG conditions. This revealed a highly reliable effect of instruction type (SG vs. DG), $\chi^2 (1, N = 90) = 6.43, p < .05, w = .27$. Further chi square analyses showed there to be no effect of triple numbers produced within either instruction types: DG - $\chi^2 (2, N = 45) = 2.22, p > .05$; SG - $\chi^2 (2, N = 45) = 4.20, p > .05$. However, collapsing the two less successful conditions (6 and 18 triples) together and comparing success rates against those achieved in the 12 triple condition confirms that in this study generating 12 triples was significantly associated with success on the task, $\chi^2 (2, N = 90) = 5.02, p < .025, w = .34$. In terms of correct rule announcements, then, the results of this study go some way toward arbitrating against the predictions of the information quantity account (Wharton et al., 1993).
4.3.2 Analysis of Triples Generated

4.3.2.1 Analysis of Triple Generation Profiles across Solvers and Non Solvers

Further analyses of the data across those solving and not solving were carried out to help arbitrate between competing accounts of DG facilitation. Recall that the triple heterogeneity theory (Vallee-Tourangeau et al., 1995) proposes that a wide search of the problem space (as indexed by the variety of triples produced) will lead to success on the task, whilst the goal complementarity theory (Wharton et al., 1993) suggests that it is the production of posvar triples which is critical for success.

Since participants were forced to produce 6, 12 or 18 triples and this may impact on the numbers of triple types produced, a proportional score of the number of each triple type generated relative to the number of triples required was calculated and used in the analysis (see Table 4.3). A series of t tests showed that with the exception of the proportion of posvar triples (which approached significance), \( t(88) = 1.98, p = .051 \), there was a significant difference between the proportions of all measures of triple heterogeneity across success. Solvers produced a lower proportion of constant positives, \( t(88) = 5.89, p < .001, d = 1.06 \), and a higher proportion of both negative triples \( t(88) = 8.124, p < .001, d = 1.31 \) and types of negative triples \( t(88) = 5.42, p < .001, d = 1.00 \). These results offer support to the triple heterogeneity account of DG facilitation, and while they do not offer direct support for the goal complementarity account it would be difficult to make the claim that they arbitrate against it.

4.3.2.2 Analysis of Triple Generation Profiles across Number of Triples Generated

Given the observation of the differing patterns of triple production across successful and non-successful participants and the pattern of performance across the requirement to produce different numbers of triples, it was decided to explore whether the requirement to produce more (or fewer) triples impacted on the types of triples produced. This analysis was conducted using the proportion of each triple type produced relative to the total number of triples elicited. Exploration of the data showed the data were not normally distributed for any of the variables; furthermore, transformation of the data did not achieve normal distributions. Kruskal Wallace one-way ANOVAs were therefore performed.

Table 4.4 shows the medians and IQR of the proportions of types of triples generated across conditions. Analysis showed that in common with the triples produced by solvers and non-solvers there were differences across all measures of triple heterogeneity with the exception of production of posvar triples, which
Experiment 2: Investigating Information Quantity Theory

Table 4-3

Proportions of Triple Types Produced across Solvers and Non-Solvers

<table>
<thead>
<tr>
<th>Triple Type</th>
<th>Success</th>
<th>N</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant positives</td>
<td>Solver</td>
<td>42</td>
<td>0.412</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>Non-solver</td>
<td>48</td>
<td>0.818</td>
<td>0.058</td>
</tr>
<tr>
<td>Posvars</td>
<td>Solver</td>
<td>42</td>
<td>0.146</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>Non-solver</td>
<td>48</td>
<td>0.085</td>
<td>0.019</td>
</tr>
<tr>
<td>Negative feedback</td>
<td>Solver</td>
<td>42</td>
<td>0.438</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>Non-solver</td>
<td>48</td>
<td>0.133</td>
<td>0.027</td>
</tr>
<tr>
<td>Number of negative</td>
<td>Solver</td>
<td>42</td>
<td>0.181</td>
<td>0.014</td>
</tr>
<tr>
<td>types</td>
<td>Non-solver</td>
<td>48</td>
<td>0.072</td>
<td>0.014</td>
</tr>
</tbody>
</table>

approached but failed to achieve significance: proportion of constant positives, $H(2) = 17.70$, $p < .001$; proportion of posvar triples, $H(2) = 5.49$, $p = .064$; proportion of negative feedback, $H(2) = 19.13$, $p < .001$; and proportion of negtypes, $H(2) = 17.70$, $p < .001$. Note that effect sizes are not reported here as although Leech and Onweguegbuzie (2002) call for their use in non-parametric tests, an effect size measure is not available for comparisons across three independent groups with data that violates the requirement for homogeneity of variance; in addition, effect size measures that assume homogeneity of variance are adversely affected when that assumption is violated.

To tease out where these differences lay Mann Whitney tests were used with a Bonferroni correction applied to all analyses leading to an alpha of .017. These showed no significant differences between the types of triples generated by those who produced 12 or 18 triples. Those producing six triples differed from those producing both 12 and 18 triples in terms of the proportion of constant positives produced; $U = 168$, $p < .001$, $N = 60$, $d = .13$ and $U = 264.5$, $p = .005$, $N = 60$, $d = .77$, with those asked to generate six triples producing a higher proportion of constant positive triples.
Those asked to generate six triples also differed from those asked to produce 12 triples in terms of both the proportion of negtypes to triples produced, $U = 256, p < .005, N = 60, d = .82$, and the proportion of triples receiving negative feedback to triples produced, $U = 169.5, p < .001, N = 60, d = 1.32$, with both more negtypes produced, and a greater proportion of triples with negative feedback being produced by participants generating 12 triples. These analyses suggest that participants asked to generate six triples failed to develop a hypothesis testing strategy. Instead, they appear to stay at the early stages of hypothesis testing behaviour in which confirmatory evidence (in this case positive instances of the hypothesised rule) is first collected before moving on to the falsification strategies as described by Tweney in his reviews of "real life" hypothesis testing amongst scientists (Tweney, 1985a, 1989).

Table 4-4

Comparison of Proportions of Triple Types Generated across Triple Number Requirements

<table>
<thead>
<tr>
<th>Proportion of triple types</th>
<th>Number of Triples Produced</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant positive</td>
<td>6</td>
<td>0.83</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.41</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.47</td>
<td>0.58</td>
</tr>
<tr>
<td>Posvar</td>
<td>6</td>
<td>0.00</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.08</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.11</td>
<td>0.22</td>
</tr>
<tr>
<td>Negative feedback</td>
<td>6</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.41</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.23</td>
<td>0.50</td>
</tr>
<tr>
<td>Negtypes</td>
<td>6</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>0.11</td>
<td>0.22</td>
</tr>
</tbody>
</table>
Chapter 4

Experiment 2: Investigating Information Quantity Theory

4.3.2.3 Comparison of Triples Produced Across Instruction Type

The triple heterogeneity account of facilitated performance on the DG version of the task evokes the observation that DG instructions promote a wider search of the problem space as indexed by the range of triples generated. To test this idea comparisons were performed on the types of triples produced by participants given SG versus DG instructions (Table 4.5). Data on all measures were skewed so Mann Whitney U tests were performed. Contrary to the predictions of the triple heterogeneity account of DG facilitation these analyses showed no differences in the types of triple generated across instruction type, although caution should be exercised when drawing conclusions since the requirement to produce specific numbers of triples may have impacted on the hypothesis testing strategies of the participants.

Table 4-5

Comparison of Proportion of Triple Types Generated across Instruction Type

<table>
<thead>
<tr>
<th>Triple Type</th>
<th>Instruction Type</th>
<th>Median</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant positive</td>
<td>SG</td>
<td>0.58</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>0.50</td>
<td>0.39</td>
</tr>
<tr>
<td>Posvars</td>
<td>SG</td>
<td>0.00</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>0.08</td>
<td>0.17</td>
</tr>
<tr>
<td>Triples with negative feedback</td>
<td>SG</td>
<td>0.33</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>0.33</td>
<td>0.38</td>
</tr>
<tr>
<td>Negtypes</td>
<td>SG</td>
<td>0.08</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>DG</td>
<td>0.16</td>
<td>0.17</td>
</tr>
</tbody>
</table>

4.3.2.4 Analysis of the Effect of Generating a Descending Triple

To explore the contrast class account discussed in the previous chapter it was necessary to examine the role of the presence of at least one descending triple on task success (Table 4.6). A chi-square analysis showed a significant association between presence of a descending triple and success on the task, $\chi^2(1, N = 90) = 29.01, p < .001, w = .57$, lending support to the notion that the presence of at least one descending triple plays a critical role in promoting rule discovery.
Table 4-6

*Frequency of Correct Rule Announcement by Presence or Absence of Descending Triple*

<table>
<thead>
<tr>
<th>Success</th>
<th>N</th>
<th>Descending Triple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solver</td>
<td>42</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>Non-solver</td>
<td>48</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>14</td>
</tr>
</tbody>
</table>

However, examination of the presence or absence of a descending triple by instruction type showed that there was no significant difference between the two groups on this measure, \( \chi^2(1, N = 90) = 1.62, p = .70 \) (see Table 4.7). This militates somewhat against the idea that DG instructions facilitate performance on the task by promoting the production of descending triples which in turn lead to success. It is important to note once again, however, that hypothesis testing strategies in the present study may well have been constrained by the requirement to produce specific numbers of triples.

Table 4-7

*Frequency of Descending Triple Generation as a Function of Instruction Type*

<table>
<thead>
<tr>
<th>Instructions</th>
<th>N</th>
<th>Descending Triple</th>
</tr>
</thead>
<tbody>
<tr>
<td>DG</td>
<td>45</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>SG</td>
<td>45</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>22</td>
</tr>
</tbody>
</table>

Given the observation that the number of triples demanded may constrain the strategies adopted and therefore success on the task it was decided to explore the absence or presence of descending triples over number of triples demanded (see Table 4.8). As those asked to produce twelve triples were differentially more successful than those required to produce either six or eighteen triples, it would be expected that if the presence of a descending triple is crucial to success participants producing twelve triples would be more likely to produce a descending triple. A chi-square analysis supported this notion, \( \chi^2(2, N = 90) = 11.43, p < .005, w = .36 \). A series of 2 x 2 chi-square tests of independence were performed to identify where this association lay,
with Bonferroni corrections applied giving an alpha value of .017. These showed that the only difference to be found was between those participants asked to generate six triples and those asked to produce twelve triples, $\chi^2 (1, N = 60) = 11.38, p < .005, w = .43$.

Table 4-8

*Frequency Counts of Presence of Descending Triple by Number of Triples Demanded*

<table>
<thead>
<tr>
<th>Number of Triples Demanded</th>
<th>N</th>
<th>Descending Triple</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Absent</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td>18</td>
<td>30</td>
<td>13</td>
</tr>
</tbody>
</table>

4.3.3 Analysis of the Effect of Generating a Posvar Triple

The goal complementarity theory (Wharton et al., 1993) predicts that it is the production of posvar triples which is key to success on the 2-4-6 task; for this reason a similar set of analyses to those just described exploring the production of negative triples was performed. These showed that the production of a posvar triple was indeed associated with success, (see Table 4.9), $\chi^2 (1, N = 90) = 4.59, p < .05, w = .22$, but that there was no association between instruction type and production of a posvar triple, $\chi^2 (1, N = 90) = 2.18, p > .05$.

Table 4-9

*Frequency of Correct Rule Announcement by Presence or Absence of Posvar Triple and Presence of Posvar Triples by Instruction Type*

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Absent</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Success</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solver</td>
<td>42</td>
<td>15</td>
<td>27</td>
</tr>
<tr>
<td>Non-solver</td>
<td>48</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td><strong>Instruction type</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SG</td>
<td>45</td>
<td>25</td>
<td>20</td>
</tr>
</tbody>
</table>
Finally, the effect of the number of triples generated on the production of a posvar triple was explored. Chi-square analysis showed there to be an effect of number of triples demanded, $\chi^2(1, N = 90) = 6.50, p < .05, w = .28$ (Table 4.10). A series of chi-square tests of independence with Bonferroni adjustments identified that there was a difference between those asked to produce 6 and 18 triples, $\chi^2(1, N = 60) = 5.41, p < .05, w = .30$, and those asked to produce 6 and 12 triples, $\chi^2(1, N = 60) = 4.28, p < .05, w = .26$. There was no difference between those asked to produce 12 and 18 triples, that is, the difference was between those required to produce 6 rather than 12 or 18 triples, which tends to indicate production of a posvar is associated with both increased triple production and successful rule discovery.

Table 4-10

<table>
<thead>
<tr>
<th>Number of Triples Demanded</th>
<th>N</th>
<th>Absent</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>30</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>30</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>18</td>
<td>30</td>
<td>11</td>
<td>19</td>
</tr>
</tbody>
</table>

4.4 Discussion

The results this study have produced some interesting findings. First, it is clear that the information quantity account (Wharton et al., 1993) does not appear to be an adequate explanation of the facilitatory effect of DG instructions in that even when the number of triples is controlled for, participants given DG instructions are more successful on the task than those given SG instructions. However, as mentioned throughout the analysis, some care must be taken in interpreting this finding as the requirement to produce a fixed number of triples may have impacted on hypothesis testing strategies. Indeed, in terms of Bayesian hypothesis testing, by explicitly requiring a particular number of triples the participants may not have reached the criterion level of confidence before being forced to announce their hypothesised rule.
Chapter 4

Experiment 2: Investigating Information Quantity Theory

It is also evident that successful rule discovery was not associated with increased levels of triple production; indeed, there was a hint that there may be an optimum number of triples that need to be produced to facilitate rule discovery. Again, this conclusion should be viewed with caution as it is possible that this observation may be an artefact of the requirement for participants to produce a specific number of triples.

It was argued in Chapter 3 that the association of improved performance with increased triple production may be due to the relaxation of the singularity principle within the hypothetical thinking framework (Evans, 2006), and the resultant institution of a Bayesian approach to the task. It was argued that this would lead to increased triple production as participants strove to achieve a threshold level of confidence in their hypotheses. On this interpretation increased triple production is an artefact of the hypothesis testing strategies adopted by participants, and the requirement to produce a specific number of triples may impact on the participants’ strategy and therefore undermine performance.

It is, however, encouraging for the 2-4-6 task as a model of scientific reasoning that the analysis of triple types produced seems to indicate that those who were required to produce only six triples overwhelmingly generated constant positive triples, that is, they appeared to be collecting confirmatory evidence for their assumed hypotheses. This reflects the hypothesis testing behaviour reported in reviews of “real world” examinations of scientific discovery in which it is reported that early tests of theories focus on confirmatory tests and then move onto disconfirmatory testing (Tweney, 1985a, 1989)

Analysis of the types of triple produced offers support for triple heterogeneity account (Valleé-Tourangeau et al., 1995) of DG facilitation in that differences were found between successful and non successful participants in all of the measures of triple variety with the exception of the proportion of posvar triples, suggesting that producing a wide range of triples does indeed promote success on the task. This same finding militates against the goal complementarity theory of DG facilitation in that contrary to predictions of this account successful reasoners did not produce more posvars than unsuccessful participants. However, further analyses reveal that production of at least one posvar is associated with success on the task which tends to indicate that production of a single instance is sufficient to alert the reasoner to the overly-restricted nature of their hypothesis.

The novel finding from Experiment 1, that DG instructions promote the production of a descending triple, which in turn promotes salience of the ascending/descending axis, is partially supported by this study. The association between production of a descending triple and success on the task shows a large effect
size (Cohen, 1988); however, the failure to find an association between instruction type and the production of a descending triple presents some difficulty for the idea of contrast class theory. Thus further work is needed to explore the notion of contrast class information playing a role in successful solving of Wason’s 2-4-6 task.

The findings of this study tend to suggest that information quantity theory (Wharton et al., 1993) in its original form is not a convincing account of DG facilitation on the Wason’s 2-4-6 task, although it is possible that this account may be subsumed within other accounts. Neither triple heterogeneity theory (Vallée-Tourangeau et al., 1995) nor the goal complementarity account (Vallée-Tourangeau et al., 1995) are fully supported, although methodological concerns centred around the demand to produce specific numbers of triples, (especially low numbers which did not seem to allow participants to develop their hypothesis testing strategies) suggest that further studies are required to investigate these accounts. A similar approach should also be taken to the observation that descending triples provide a cue to the salient ascending/descending dimension of the rule, and future studies should investigate this contrast class account of the effect both by post hoc analyses and direct manipulation.
Experiment 3: Investigating Goal Complementarity Theory

Chapter 5

Experiment 3: The Goal Complementarity Theory of Dual Goal Facilitation

5.1 Introduction

Experiments 1 and 2 undermine both the positivity bias (Evans, 1983; 1989) and information quantity (Wharton et al., 1993) accounts of DG facilitation. However, there is some supporting evidence for both the goal complementarity (Wharton et al., 1993) and the triple heterogeneity (Vallée-Tourangeau et al., 1995) accounts of the DG effect. In addition, analyses have been presented which suggest that contrast class information may have a role to play in successful performance on the DG task. It was therefore decided to investigate these competing accounts of DG facilitation, with a particular focus on the most intuitively appealing account, that is, goal complementarity theory. As with Experiments 1 and 2, post hoc analyses will also consider the evidence for other accounts of DG facilitation.

Recall that Wharton et al.'s (1993) goal complementarity theory relies on three assumptions: (1) the embedding of initial hypotheses within the target rule (Wetherick, 1962); (2) the complementary nature of the DAX and MED rules (triples can only be DAX or MED); and (3) the tendency for people to adopt a “positive test strategy” on the 2-4-6 task (Klayman & Ha, 1987, 1989), which involves generating triples that match a current hypothesis. Based on these assumptions, Wharton et al. propose that facilitation arises because a positive test of an overly broad MED hypothesis (using a triple like “2-3-10”) can lead to DAX feedback, which then promotes the realisation that the current DAX hypothesis is too narrow.

Although the first assumption of Wharton et al.'s framework is indisputable, and the third assumption has received considerable support, the second assumption—that goal complementarity is essential for facilitation—has led to contradictory findings. In Tweney et al.'s (1980) original DG study the relationship between DAX and MED rules was ambiguous, yet success was high. Wharton et al. (1993), however, showed that when the complementary nature of the DAX and MED rules was not made explicit then DG instructions failed to facilitate performance. In contrast, Vallée-Tourangeau et al. (1995) claimed to have undermined goal complementarity theory by demonstrating DG facilitation even when participants were explicitly given non-complementary DAX and MED rules. On closer inspection, however, Vallée-
Tourangeau et al.'s evidence seems fragile. This is because although participants were told that triples could be "DAX, MED, or neither" (and, in another condition, "DAX, MED, neither or both"), it remained the case that standard (complementary) DAX and MED rules and feedback were always used. Thus, unbeknownst to participants, the logical relationship between rules remained strictly complementary and they never received feedback other than DAX or MED. As such, Vallée-Tourangeau et al.'s methodology contained flaws that could well have produced artefactual evidence opposing goal complementarity theory. The present experiment tackled such flaws through a more effective manipulation of the presence versus absence of goal complementarity.

Notwithstanding the limitations of Vallée-Tourangeau et al.'s (1995) results there is other evidence that undermines the viability of the goal complementarity account of DG facilitation. As noted above, goal complementarity theory claims that it is the positive testing of overly broad MED hypotheses (using triples like "2–3–10")—and the surprising DAX feedback that results—which then leads to an appreciation that the DAX hypothesis needs broadening. The theory therefore predicts that it is the production of a "posvar triple" (an ascending triple that does not increase by equal intervals) that is essential for task success. Experiments 1 and 2, however, have demonstrated that it is the production of "descending triples" (e.g., "6–4–2" or "9–4–3") rather than posvar triples that is the dominant predictor of task success. This effect could be explained by the suggestion that the production of a descending triple provides a strong class cue to the likely scope of the DAX rule (i.e., that it relates to "ascending" sequences as opposed to "descending" ones). Further work is required, however, both to replicate these findings and to flesh out the details of the contrast class theory of triple testing in the DG paradigm.

In the case of triple heterogeneity theory, Vallée-Tourangeau et al. (1995) have demonstrated that DG instructions foster more creative triple exploration in terms of the variety of triples that are tested compared with standard instructions. It remains unclear, however, why the mere breadth of triples examined should enhance rule discovery. As such, triple heterogeneity seems more descriptive than explanatory. In addition, direct testing of the theory has not been undertaken, possibly because it is difficult to imagine a manipulation that would allow such direct testing. For this reason post hoc analyses will once again include an assessment of the range of triples produced in order to provide further evidence pertaining to this theory.

Table 5.1 lists the three main theories of DG facilitation that were examined in this experiment along with associated predictions relating to the triple-generation profiles expected to mediate task success. The crucial test of goal complementarity...
theory in the present experiment involved manipulating the relationship between the DAX and MED rules such that in two DG conditions DAX and MED were no longer perfect complements. Standard SG and DG tasks were retained as control conditions and triple generation was kept constant at 10 triples. Although the results of Experiment 2 tended to undermine information quantity theory, in the present experiment triple generation was kept constant as a way of controlling for any influence of triple number. The decision to impose a 10-triple limit derived from two previous studies (Klayman & Ha, 1989; Vallée-Tourangeau et al., 1995) that used a similar constraint as an effective way to standardise any influence of information quantity on rule discovery, whilst still affording participants an opportunity to examine a variety of triple types should they wish to. Vallée-Tourangeau et al. (1995) note that 10 triples seems a particularly good cut-off since a majority of participants are eager to announce their hypotheses just before this point is reached. As such, this criterion is likely to be minimally disruptive to participants’ natural triple testing behaviour whilst permitting tight controls over the quantity of triples tested across conditions.

5.2 Method

5.2.1 Participants

Eighty students at sixth-form colleges local to the University of Derby voluntarily took part in the study. No participants had received training in logic or reasoning.

5.2.2 Design

An independent measures design was employed. This entailed a standard single goal (SG) condition in which participants were asked to discover the DAX rule, with feedback for generated triples in the form of DAX or MED. There were three dual goal (DG) conditions in which participants were asked to discover both the DAX and the MED rules. For DG-2 a standard DG task was used that entailed a complementary goal structure and DAX or MED feedback. DG-3 was a non-complementary condition where feedback could be DAX, MED or “neither DAX nor MED”. DG-4 was another non-complementary condition, where feedback could be DAX, MED, “neither DAX nor MED” or “both DAX and MED”.

- 72 -
Table 5-1

*Theories of Dual Goal Facilitation, Including Predictions and Empirical Tests*

<table>
<thead>
<tr>
<th>Account</th>
<th>Features</th>
<th>To Test</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal complementarity theory (Wharton et al., 1993)</td>
<td>Only complementary goals cause DG facilitation.</td>
<td>Manipulate complementarity of DAX and MED rules</td>
<td>DG facilitation not evident when non-complementary rules used</td>
</tr>
<tr>
<td></td>
<td>Task success dependent on the production of at least one posvar triple</td>
<td>Compare presence of posvars across conditions as well as for solvers versus non-solvers</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Variety of triples produced mediated between type of instruction and success on the task</td>
<td></td>
</tr>
<tr>
<td>Triple heterogeneity theory (Vallée-Tourangeau, Austin, &amp; Rankin, 1995)</td>
<td>DG facilitation depends upon the production of a wider variety of triple types</td>
<td>Analyse the variety of triple types across conditions as well as for solvers versus non-solvers</td>
<td>Variety of triples produced mediated between type of instruction and success on the task</td>
</tr>
<tr>
<td>Contrast class theory (Experiment 1)</td>
<td>DG instructions promote production of descending triple</td>
<td>Compare presence of descending triple across instruction mode and for solvers versus non-solvers</td>
<td>Production of a descending triple mediated between instruction mode and success on the task</td>
</tr>
</tbody>
</table>
5.2.3 Materials and Procedure

Participants were tested in groups of two to four, with each participant producing written triples and receiving written feedback. Participants were provided with a sheet on which to record triples, feedback and their best guess at the rule after producing 10 triples. As with Experiments 1 and 2 participants were informed that they would be permitted only a single attempt at stating the rule.

Participants in the SG condition received instructions relating to a unique to-be-discovered rule: “I have in mind a rule that specifies how to make up sequences of three numbers (triples), and your task is to discover this rule. Triples that fit my rule are called DAX triples and other triples are called MED triples. To start you off, I can tell you that 2-1-6 is a DAX triple. In order to discover my rule you should produce further number triples, and I will tell you whether they are DAX triples, or whether they are MED triples”. Participants in other conditions were instructed to discover two rules, one called DAX, the other called MED. In all conditions the DAX rule was “any ascending sequence”. In DG-2 the MED rule was “any other sequence” (logically complementary to DAX), in DG-3 the MED rule was the non-complementary “descending triples”, whilst in DG-4 the MED rule was the non-complementary “odd numbers”.

5.3 Results and Discussion

5.3.1 Task Success across Conditions

As with previous DG studies only participants’ DAX rule statements were considered when determining task success. The data presented in Table 5.2 for “total solvers” and “total non-solvers” indicate that participants in the DG-4 condition had the most difficulty with the task with only 15% correctly announcing the rule, which reflected worse performance than in the standard SG condition where there were 35% solvers. One reason for the poor performance in the DG-4 condition may relate to the nature of presented feedback, where triples could be both DAX and MED. This feedback certainly appeared to confuse participants, perhaps hindering their understanding of the task requirements. There are, however, deeper conceptual reasons for the difficulty of the DG-4 condition which relate to the theoretical assumptions of the contrast class theory. These issues are discussed below when examining the mediating role of descending triple production for task success.

In contrast to the poor performance seen under DG-4 and SG instructions, the number of participants discovering the correct DAX rule in the DG-3 condition (75%) was comparable (indeed marginally superior) to the standard DG-2 condition (65%, see Table 5.2). A chi-square test of independence indicated that these differences in solvers versus non-
solvers across conditions were reliable, $\chi^2(3, N = 80) = 18.24, p < .001, w = .48$. The high success rate in the non-complementary DG–3 condition does not lend support to Wharton et al.’s (1993) goal complementarity theory.

Table 5-2

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Solvers</th>
<th>Non-solvers</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>20</td>
<td>7 (35%)</td>
<td>13 (65%)</td>
</tr>
<tr>
<td>DG–2</td>
<td>20</td>
<td>13 (65%)</td>
<td>7 (35%)</td>
</tr>
<tr>
<td>DG–3</td>
<td>20</td>
<td>15 (75%)</td>
<td>5 (25%)</td>
</tr>
<tr>
<td>DG–4</td>
<td>20</td>
<td>3 (15%)</td>
<td>17 (85%)</td>
</tr>
</tbody>
</table>

Note: Percentages in parentheses.

5.3.2 Triple Types across Condition

To code participants’ triple generation profiles an extension of the triple-coding system introduced by Vallée-Tourangeau et al. (1995) was used. According to this scheme a “negtype score” can be denoted by an integer between 0 and 8, and is distinct from the total number of generated triples receiving negative feedback. One further index of triple variety was also computed: the mean number of all types of triple produced, whether positive or negative. This index, which ranged from 1 to 10, is referred to as the “triptype score” and could be argued to be a truer reflection of the variety of hypotheses considered than the more restricted negtype index.

According to the triple heterogeneity theory of DG facilitation, DG instructions foster success by encouraging a wide exploration of the problem space, as indexed by the variety of triple types produced. To investigate this notion further a series of one way ANOVAs was pursued on the types of triples produced across conditions. Table 5.3 summarises participants’ triple-generation profiles across conditions for the indices of triple heterogeneity. A significant effect of condition was found on the number of triples receiving negative feedback, $F(3, 76) = 3.62, p = .017, \eta^2 = .13$, and the number of constant positive triples, $F(3, 76) = 3.38, p = .023, \eta^2 = .12$. No other effects were found. Post hoc Bonferroni tests indicated that the only reliable differences were between the number of triples with negative feedback in the SG versus DG–2 conditions ($p = .04$), and the number of constant positive triples in the SG versus DG–2 conditions ($p = .05$). These analyses show that although the standard DG–2 instructions increase production of negative triples and decrease production of constant positives relative to SG instructions there is no increased range of triple production under any DG instructions,
Table 5-3

**Participants’ Triple-Generation Profile across Conditions**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Constant Positives</th>
<th>Posvars</th>
<th>Negtype Score</th>
<th>Negative Feedback</th>
<th>Triptype Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
</tr>
<tr>
<td>SG</td>
<td>6.75</td>
<td>.78</td>
<td>1.15</td>
<td>.41</td>
<td>1.05</td>
</tr>
<tr>
<td>DG–2</td>
<td>4.40</td>
<td>.48</td>
<td>1.80</td>
<td>.36</td>
<td>2.00</td>
</tr>
<tr>
<td>DG–3</td>
<td>4.50</td>
<td>.58</td>
<td>1.75</td>
<td>.52</td>
<td>1.95</td>
</tr>
<tr>
<td>DG–4</td>
<td>4.60</td>
<td>.58</td>
<td>2.85</td>
<td>.43</td>
<td>1.55</td>
</tr>
</tbody>
</table>

*Note:* SE = standard error. Triple-generation profiles in terms of mean number of constant positive triples, mean number of posvar triples, mean negtype score (ranging from 0 to 8), mean number of triples receiving negative feedback, and triptype score (ranging from 1 to 10).
despite high rates of task success in the DG-2 and DG-3 conditions. These results undermine the notion that DG instructions lead to success by promoting more creative triple testing. Likewise, the failure to find reliable effects across conditions relating to the generation of posvar triples contradicts the supposed importance of such triples for task success claimed by goal complementarity theorists.

5.3.3 Triple Generation Strategies Associated with Task Success

To elucidate the triple-generation strategies which lead to rule discovery it was necessary to compare the triple production profiles of solvers versus non-solvers, collapsing data across conditions (Table 5.4). A series of t-tests revealed that solvers produced fewer constant positives than non-solvers, \( t(78) = 2.98, p = .004, d = .64 \), and more triples receiving negative feedback, \( t(78) = 4.61, p < .001, d = .66 \). Solvers also showed larger negtype scores, \( t(78) = 2.19, p = .031, d = .48 \), larger triptype scores, \( t(78) = 2.19, p = .031, d = .48 \), and produced more descending triples, \( t(78) = 6.02, p < .001, d = 1.12 \). These analyses lend some support to the idea that people who solve the 2-4-6 task exhibit rather different triple-generation strategies than non-solvers, producing a greater range of triple types as well as more triples receiving negative feedback.

Table 5-4

<table>
<thead>
<tr>
<th>Measure of Heterogeneity</th>
<th>Solvers</th>
<th>Non-Solvers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
</tr>
<tr>
<td>Constant Positives</td>
<td>4.11</td>
<td>0.34</td>
</tr>
<tr>
<td>Posvars</td>
<td>1.76</td>
<td>0.33</td>
</tr>
<tr>
<td>Negtype Score</td>
<td>1.97</td>
<td>0.19</td>
</tr>
<tr>
<td>Negative Feedback</td>
<td>4.24</td>
<td>0.33</td>
</tr>
<tr>
<td>Triptype Score</td>
<td>3.63</td>
<td>0.20</td>
</tr>
<tr>
<td>Descending Triples</td>
<td>2.50</td>
<td>0.28</td>
</tr>
</tbody>
</table>

*Note: SE = standard error. Triple-generation profiles in terms of mean number of constant positive triples, mean number of posvar triples, mean negtype score (ranging from 0 to 8), mean number of triples receiving negative feedback, and triptype score (ranging from 1 to 10).*
Finally, the contrast class account of DG facilitation was examined. This theory proposes that task success is not determined by the range of triples that are generated (as espoused by triple heterogeneity theory) but rather by the production of at least one descending triple (i.e., where \( a > b > c \)) as part of this varied triple exploration. Under the contrast class theory the generation of a descending sequence is viewed as being a necessary condition for task success, since receiving MED feedback for such a triple is vital for cueing the reasoner toward the realisation that the MED versus DAX contrast class falls on the dimension “descending versus ascending”. To support the necessity of descending triple production for task success two pieces of evidence are essential. First, it is critical to demonstrate that task success is very likely given the production of at least one descending triple. Second, it must be shown that task success is very unlikely given the generation of at least one triple that is neither descending nor ascending (NB: “ascending” triples are non-diagnostic in that all participants are given the ascending 2-4-6 sequence as their seed triple).

To examine the first prediction a chi square test of independence on the presence versus absence of a descending triple across conditions (Table 5.5) was conducted. This revealed a reliable effect, \( \chi^2 (3, N= 80) = 9.02, p = .029, \omega = .34 \), with the lowest levels of descending triple generation arising in non-facilitatory SG condition and the highest levels arising in facilitatory conditions (DG–2 and DG–3). The DG–4 condition presents an interesting counterpoint to this general pattern, since around 50% of participants in this condition generated a descending triple, yet task success was very limited. However, it is possible to interpret this finding in line with contrast class theory, which claims that to solve the task people first need to generate the “MED equals descending” hypothesis before they can then be cued toward the realisation that the MED versus DAX contrast class potentially falls on the dimension “descending versus ascending”. In the DG–4 condition the underlying MED rule was actually “odd numbers” such that there was little or no opportunity for participants to receive the cue that the MED rule might relate to descending triples. Indeed, most of the MED feedback that participants received would have had little to do with the descending nature of triples. It is thus unsurprising that descending triples were generated but task success was low.
Chapter 5

Experiment 3: Investigating Goal Complementarity Theory

Table 5-5

Frequency of Participants Producing at Least one Descending Triple across Conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Absent</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>20</td>
<td>11 (55%)</td>
<td>9 (45%)</td>
</tr>
<tr>
<td>DG-2</td>
<td>20</td>
<td>3 (15%)</td>
<td>17 (85%)</td>
</tr>
<tr>
<td>DG-3</td>
<td>20</td>
<td>4 (20%)</td>
<td>16 (80%)</td>
</tr>
<tr>
<td>DG-4</td>
<td>20</td>
<td>7 (35%)</td>
<td>13 (65%)</td>
</tr>
</tbody>
</table>

Note: Percentages in parentheses.

To examine the prediction that no other types of triple (apart from descending ones) have the capacity to promote task success, a second chi square test of independence was conducted on the dataset depicted in Table 5.6, which crossed instructions with production of triples of these types. This showed that instruction type was not associated with the production of triples that were neither ascending or descending, $\chi^2 (3, N=80) = 7.27, p = .064$ In sum, these findings attest to the importance of descending triple production in mediating task success, as predicted by the contrast class account.

Table 5-6

Presence versus Absence of at Least One Triple that was neither Descending nor Ascending across Conditions

<table>
<thead>
<tr>
<th>Task Success</th>
<th>N</th>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>SG</td>
<td>20</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>DG-2</td>
<td>20</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>DG-3</td>
<td>20</td>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>DG-4</td>
<td>20</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>
5.3.5 Discriminating between the Goal Complementarity and Contrast Class Information Accounts

The goal complementarity theory emphasises that production of posvar triples is instrumental in facilitating success, whereas the contrast class theory dismisses the importance of such triples, instead proposing that production of at least one descending triple is the critical determinant for rule discovery. To clarify the relative roles of posvar and descending triples for task success the dataset was modelled using logistic regression. An initial analysis in which the dichotomous predictor variable of descending triple was present versus absent was regressed onto solution success was highly significant, $B = 2.55$, Wald = 14.24, $p < .001$. Indeed, participants generating a descending triple were twelve times more likely to discover the DAX rule than those who didn't generate a descending triple. A second model in which presence versus absence of a posvar triple was regressed onto solution success failed to reach significance, $B = .51$, Wald = 1.16, $p > .05$. This finding undermines a key prediction of goal complementarity theory that it is posvar generation that is critical for success on the task.

5.3.6 Path Analysis

The final examination of the dataset involved conducting a path analysis (Baron & Kenny, 1986) to establish whether producing at least one descending triple played a mediating role in task success (intervening between task instruction and rule discovery). In preparation for this analysis data from the DG–2 and DG–3 conditions were collapsed since the triple-production profiles of these participants were very similar. The DG–4 data was also excluded because of the difficulty that participants had found with this task. Since the quantitative logic of logistic regression does not allow for path analysis it was necessary to use linear regression. As noted previously this is not ideal given the dichotomous nature of the relevant variables. The $p$ values obtained from calculations based on logistic and linear regression are therefore presented to assess the viability of this procedure (Table 5.7). Figure 5.1 shows the path diagram obtained when the mediating role of producing at least one descending triple was examined to determine its effect on success. A Sobel test indicated that the mediating effect was significant, Goodman (1) = 2.62, $p < .01$. 
Chapter 5

Experiment 3: Investigating Goal Complementarity Theory

Table 5-7

Summary of Mediator Analysis of the Relationship between Goal, Production of at Least One Descending Triple, and Success, Using Linear Regression

<table>
<thead>
<tr>
<th>Variables in Regression Analysis</th>
<th>p values obtained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear Regression</td>
</tr>
<tr>
<td>Goal onto success</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Goal onto descending triple</td>
<td>.009</td>
</tr>
<tr>
<td>Goal and descending triple onto success</td>
<td>Goal</td>
</tr>
<tr>
<td></td>
<td>Descending triple</td>
</tr>
</tbody>
</table>

Note: Goal is dichotomous (SG vs. DG); Descending triple is dichotomous (present vs. absent); Success is dichotomous (solver vs. non-solvers)

Figure 5-1

Path Diagram Showing the Mediating Effect of Producing at Least One Descending Triple on Task Success

5.4 General Discussion

This study was designed to arbitrate between theories of facilitation in DG variants of Wason's rule discovery paradigm using instructional manipulations that altered the relationship between the to-be-discovered rules. Wharton et al.'s (1993) goal complementarity theory is challenged by these results in two key respects. First, when DAX and MED rules were no longer logically complementary (some triples fitted neither rule) performance on this DG version remained high—which contradicts the theory's central assumption that the complementary nature of DAX and MED rules is critical for facilitated performance. Second, the theory assumes it is production of posvar triples that promotes rule discovery, yet the logistic regression demonstrated this was not the case.
Chapter 5

Experiment 3: Investigating Goal Complementarity Theory

It must be acknowledged, however, that the experimental design contained a confound in that the conditions differed not only in terms of the relation between the DAX and MED rules, but also in the definition of MED (e.g., it denoted “descending” triples in one condition and “odd numbers” in another), whereas the definition of DAX remained constant across conditions. One way to disentangle the effects of the relation between the rules and actual rule content would be to implement a second set of conditions where the MED rule remains constant but where DAX changes across conditions. A successful replication of the present findings would be predicted in this case given the considerable body of evidence to support the assumption that the precise definitions of the target rules do not differentially affect solution rates (e.g., Gorman et al., 1987; Klayman and Ha, 1989; Poletiek, 1996; Rossi, Caverni, & Girotto, 2001). In Rossi et al.’s (2001) study, for example, the target rule was “descending triples” yet rule discovery rates (16%) were comparable to those observed for the standard “ascending triples” rule. Given the evidence in the literature, it is therefore unlikely that the results are explicable in terms of a rule-content confound rather than in terms of the rule complementarity manipulation.

In terms of triple heterogeneity theory (Vallée-Tourangeau et al., 1995) the analysis of triple-generation profiles across conditions provided little support for the idea that people pursue broader explorations of the problem space under DG instructions, with no measures of triple variety revealing reliable differences. At the same time, triple variety did seem to be implicated in task success when triple-generation profiles were collapsed across conditions, since solvers demonstrated larger negtype and triptype scores. However, the lack of any significant differences in indices of triple variety across conditions remains difficult to reconcile with a simple triple-heterogeneity explanation of DG facilitation, although this account may be subsumed within other accounts of DG facilitation. For example, it is possible that the generation of a wide range of triples is instrumental in producing a contrast class triple which in turn leads to facilitated performance on the task (see below).

The final theory of interest—the contrast class theory—achieved a good degree of support from the observation in Experiment 3 that production of at least one descending triple of the form \(a > b > c\) was highly associated with task success, whereas production of triples other than descending ones was a weaker predictor of rule discovery. In addition, since the contrast class theory views neither goal complementarity nor the generation of posvar triples as relevant factors for task success, this theory can accommodate the observations that non-complementary DG-3 instructions promoted high solution rates yet posvar triples were not related to success. In addition to undermining goal complementarity theory both the high rate of success in the non complementary DG-3 condition and the poor performance in the
DG-4 conditions can be readily explained by the contrast-class theory. In the DG–3 condition the MED rule provides an explicit contrast class cue to the “ascending” dimension of the DAX rule; indeed the cue provided by the MED rule in this condition is stronger than that provided in the standard DG condition in which “descendingness” is only one of a number of possible relational rules which are represented by MED feedback. On the other hand, in the DG–4 condition the MED rule does not specify opposite contrast sets, rather it focuses on non-relevant “object properties” (Cherubini et el., 2005), that is, attributes of numbers in the generated triple, rather than relational regularities which are concerned with comparing relationships across triples. This focus on object regularities means that rather than the MED rule providing a cue to the scope of the to-be-discovered DAX rule it instead provides a distraction as participants seek two unrelated rules. It is perhaps unsurprising, therefore, that performance in this condition was the poorest.

The observations arising from Experiment 3 led to a decision to pursue a small-scale and highly focussed follow-up study to explore whether the findings could be generalised to a DAX rule other than the standard “ascending numbers” one. In order to allow comparability across studies it was decided to vary the DAX rule in line with the MED rules used in Experiment 3. Table 5.8 summarises the DAX rules and associated MED rules across conditions, as well as the presented exemplar triple for each DAX rule. It can be seen from the table that when seeking DAX rules of descending numbers and odd numbers the standard “ascending numbers” is the MED rule in the DG–3 and DG–4 conditions respectively, thus providing some measure of comparability with Experiment 3.

A further 149 students (all aged over 18 years old) at sixth-form colleges local to the University of Derby voluntarily took part in the study. A 2 (rule) x 4 (relationship of rules) fully independent measures design was employed. The two rules used were “any descending sequence” and “three odd numbers”. In the standard SG conditions participants were asked to discover the DAX rule, with feedback for generated triples in the form of DAX or MED. For each rule there were three DG conditions in which participants were asked to discover both the DAX and the MED rules. For DG–2 a standard DG task was used that entailed a complementary goal structure and DAX or MED feedback. DG–3 was a non-complementary condition where feedback could be DAX, MED or “neither DAX nor MED”. DG–4 was a second non-complementary condition, where feedback could be DAX, MED, “neither DAX nor MED” or “both DAX and MED” (see Table 5.8). The procedure was identical to that employed in Experiment 3 with the exception that both the target DAX rule and the associated exemplar were varied across conditions.
Experiment 3: Investigating Goal Complementarity Theory

Table 5-8

Definitions of DAX and MED Rules across Experiment 3 and the Present Experiment

<table>
<thead>
<tr>
<th>DAX Rule and exemplar</th>
<th>MED Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascending 2-4-6</td>
<td>Any non ascending triple</td>
</tr>
<tr>
<td>Descending 6-4-2</td>
<td>Any non descending triple</td>
</tr>
<tr>
<td>Odd numbers 3-3-3</td>
<td>Any triple containing an even number</td>
</tr>
</tbody>
</table>

The purpose of the present experiment was to verify that varying the to-be-discovered rule did not affect the rule discovery performance of participants in standard SG, DG or non-complementary DG conditions. Based on the success rates achieved in Experiment 3 it was expected that success rates would be highest in the facilitatory DG-2 and DG-3 conditions, and poorest in the DG-4 condition. Chi square tests of independence were performed on the data for each rule and these showed that this pattern of performance was evident for both of the rules used in this experiment, $\chi^2 (3) = 25.68, N = 77, p < .001$, $w = .58$, and $\chi^2 (3) = 19.59, N = 77, p < .001$, $w = .50$, for the “descending numbers” and “three odd numbers” rules respectively (see Table 5.9). As such, it is clear that DG facilitation in the absence of logical complementarity is a generalisable phenomenon. Furthermore, the pattern of performance in this study is in line with the predictions of the contrast class theory which proposes that facilitation will be greatest when the MED rule provides clear contrast class information. The MED rule is simply not useful in providing contrast class information in the DG-4 condition for either rule; in the “odd numbers” rule the DG-4 MED rule is “ascending sequences”, whereas in the descending numbers rule the MED rule is “odd numbers”. On the other hand it is apparent for all three DAX rules explored that it is the DG-3 condition which provides the best contrast class cues in terms of the definition of MED; indeed, for the “odd numbers” DAX rule the
Chapter 5

Experiment 3: Investigating Goal Complementarity Theory

SG and DG-2 MED rules are not really that helpful in terms of contrast class cues, so it is therefore not surprising that DG-3 leads to the best facilitation in this condition.

Table 5-9

<table>
<thead>
<tr>
<th>Condition</th>
<th>Descending sequence</th>
<th>Odd numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Successful</td>
</tr>
<tr>
<td>SG</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td>DG-2</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>DG-3</td>
<td>16</td>
<td>88</td>
</tr>
<tr>
<td>DG-4</td>
<td>20</td>
<td>15</td>
</tr>
</tbody>
</table>

The patterns of performance in both this experiment and Experiment 3 appear to undermine the goal complementarity explanation of DG facilitation by directly manipulating the logical relationship between the two rules and, instead, seem to provide some support for the contrast class account. Other accounts of DG facilitation rely more heavily on post hoc analyses of triple generation profiles of participants across both success and instruction type. Triple heterogeneity theory (Vallée-Tourangeau et al., 1995) proposes that it is the range of triples generated which is instrumental in facilitating performance using DG instructions, thus according to this theory DG instructions should elicit a greater range of triples tested, and that solvers will produce a wider range of triples than non solvers. The variation of the DAX rule in the present study meant that the definition of types of triples would need to varied; thus for the “descending sequence” rule a constant positive triple would become a triple which descends in equal increments, and a posvar triple one that descends with unequal intervals. However, the definition of a constant positive and a posvar in the “odd numbers” rule is far more problematic. One possibility would be to elicit the types of hypotheses generated by participants in the SG condition and use these as a basis for the coding of triples; thus a constant positive triple would be one which was an example of an over-restricted DAX hypothesis, and a posvar one which relaxed that restriction. A review of the incorrect rule announcements in the SG condition showed that eleven of the fourteen incorrect rules were indeed over-restricted forms of
an “odd numbers” rule, however, there were four different restrictions on the rule (repetition, prime numbers, multiples of three, ending in three). It was therefore decided that it would be inadvisable to pursue this method for defining a posvar for this rule and therefore any analyses relating to posvar generation.

The negtype and triptype measures of heterogeneity are also somewhat problematic when using the two rules in this present experiment. In terms of the “odd numbers” target rule a negative triple is one which contains at least one even number; the relational coding of negative triples used thus far is therefore redundant. Whilst it was recognised that the negtype measure for the “descending sequences” rule is conceptually equivalent to the negtype measure for the standard “ascending numbers” rule (with the exception of coding constant positives as descending in equal intervals, and posvars as descending with non-equal intervals) it was decided that the difficulties associated with triple coding in the “odd numbers” condition prevented the use of negtypes as a meaningful measure of triple heterogeneity. A similar position is taken with the “negative feedback” measure in the present experiment. Definitional issues relating to the coding of triples in the “odd numbers” rule therefore prevented post hoc analyses of triple generation profiles, and therefore no evidence regarding triple heterogeneity theory or contrast class account of DG facilitation can be gained from these data.

Overall, then, the two experiments reported in this chapter appear to provide some good support for the assumptions of the contrast class theory of DG facilitation. How, though, can the evidence for the contrast class account be placed within a broader theoretical framework that describes the processes that underpin DG facilitation? To address this question it is first necessary to tackle the issue of how a descending triple can lead to rule discovery and also to consider the mechanism by which DG instructions encourage production of a descending triple in the first place. DG findings may best be accommodated within the general “iterative counterfactual model” (ICM) of hypothesis testing developed by Oaksford and Chater (1994), an extension of Farris and Revlin’s (1989a, 1989b) counterfactual model (see Fig 5.2).

The ICM can readily be illustrated by an example that starts with the participant being given the 2-4-6 triple as an instance of the experimenter’s rule. Cherubini et al. (2005) have shown that this seed triple encourages the formulation of a hypothesis that maintains the maximum amount of relational information; we could, therefore, assume that the participant will establish an initial hypothesis (H) of “even numbers ascending by 2”. According to the ICM the participant would test this H by
developing an alternative hypothesis (H'), which would vary from the original H on only one axis. An example H' would be "odd numbers ascending by 2", which varies on the odd/even axis but which retains "ascending numbers" and "intervals of 2". A triple that matches H' would then be generated (e.g., 3-5-7), which would receive positive feedback. The participant adds this triple to the list of positive instances of their current H, and searches for common properties of triples within this list to derive a new H, which would then be tested counterfactually via a new H'.

By this method H is continually expanded and adjusted to fit new information such that the ICM explains both the context of justification (testing triples that match the new H') and the context of discovery (the generation of a new H). Contrast class theory - which claims that task success is dependent on the production of an a > b > c triple that matches the H' "descending numbers" - maps readily onto the ICM. A descending triple differs from the previous 3-5-7 example in that it receives negative feedback. Rather than increasing the list of positive instances of H the participant now has a triple that does not conform. Furthermore, comparison of triples which conform to H against those which do not will show that one axis in which they differ is that of ascending/descending. While it is true that participants may suggest a triple that varies on more than one axis, iterative counterfactual testing is concerned with exploring one common feature at a time; it is, therefore, reasonable to suppose that when comparing examples that fit or do not fit a hypothesis participants will focus on one axis on which triples contrast.

This tendency to focus on one axis at a time also seems to concur with Vallée-Tourangeau and Payton's (2008) findings showing that SG success is dramatically improved by deploying external, graphical representations of triple properties. One explanation of this finding is that external representations lessen demands on limited working memory resources. It may be, then, that working memory constraints are the root cause of participants considering and testing single common features when examining the list of positive triples, or single contrasting features when comparing positive and negative triples.

Having discussed a possible mechanism linking descending triple production to task success the issue of why DG instructions promote the generation of descending triples is now addressed. One mechanism might be that under DG instructions participants first attempt to find the DAX rule (i.e., the first phase of DG testing is identical to SG testing). From this it could be assumed that participants will produce an overly-restricted DAX rule such as "numbers ascending by 2", and, having seemingly asatisfied the requirement to discover DAX, will move on to discovering MED. Unlike DAX participants have no instance of the MED to inform initial
hypotheses, but they do know that DAX and MED are different and that DAX seems to be "numbers ascending by 2". It seems reasonable to assume that participants will use this information to generate their initial MED hypothesis so that it differs from their DAX hypothesis as much as possible (i.e., it is the "opposite" of DAX). It is easier to produce the opposite of "ascending" (which logically opposes "descending") than of "by 2" (where the opposing set of "by all numbers other than 2" is fuzzy and confusing). It could be contended that this is the mechanism through which the requirement to search for the MED rule triggers the generation of a descending triple. Whilst this mechanism explains the production of an all-important descending triple it does not explain how this facilitates the generalisation of the overly-restricted "numbers ascending by 2" DAX hypothesis, and further investigation is required to explore this issue. For example, no research appears to have examined participants' stated MED rules, although it has been suggested that people formulate MED rules that are a subset of DAX, such as "ascending odd numbers" (Rossi, Caverni, & Girotto, 2001).

Results of this study have helped to advance the theoretical understanding of DG facilitation in Wason's 2-4-6 task. The study demonstrates that complementary rules are not essential for task success, which indicates that goal complementarity theory provides an inadequate account of DG facilitation, although it was noted that a possible confound may have been introduced by varying the definition of the MED hypothesis. For this reason it would be prudent to carry out a follow up study in which the definition of the DAX rule is systematically varied in line with the MED rules used in this study to explore what effect this may have on hypothesis testing behaviours. While the analyses of participants' triple-generation profiles do provide some support for triple heterogeneity theory of DG facilitation as being linked to more varied and creative triple testing, it is suggested that a more convincing account of DG effects—including the vital role of descending triples in enhancing rule discovery—is provided by the contrast class theory, which itself maps well onto the broader hypothesis testing framework offered by Oaksford and Chater's (1994) iterative counterfactual model. However this model needs to be extended to account fully for DG instructions and additional empirical work is required to clarify the mechanism by which the production of a descending triple leads to the generalisation of overly-restrictive hypotheses.
Chapter 6

Experiment 4: Extending the Contrast Class Theory of Dual Goal Facilitation to Different DAX Rules

6.1 Introduction

The work presented in the thesis so far has explored contemporary accounts of DG facilitation in Wason's rule discovery paradigm and has demonstrated that no current theory appears to provide a fully adequate explanation of the DG effect. It is acknowledged, however, that evidence against the triple heterogeneity theory (Vallée-Tourangeau et al., 1995) and the information quantity theory (Wharton et al., 1993) is less convincing than evidence that is contrary to the positivity bias theory (Evans, 1989) and the goal complementarity theory (Wharton et al., 1993). A new account of the DG facilitation effect has been developed which suggests that it is the production of a contrast class triple which is most associated with success on the task, and it has been argued that this contrast class account can be accommodated within Oaksford and Chater's (1994) iterative counterfactual model (ICM). However, with the exception of the study that was briefly reported in the discussion of Experiment 3 - which tested the generality of the evidence against the goal complementarity theory - the majority of research reported in the thesis has only been concerned with Wason's (1960) standard “ascending numbers” rule. It was decided, therefore, that before developing the contrast class account more fully it would be prudent to investigate whether the presence of a contrast class triple can also be demonstrated to be associated with task success in the DG paradigm when other DAX rules are employed. One candidate rule that has been reported in the literature as an alternative to Wason's traditional “ascending numbers” is Gorman's (1987) “three different numbers” rule. Despite the difficulty that participants have in discovering this rule - which is even more general than Wason's original rule - Gorman reported a DG facilitation effect, thus making it highly suitable for use in the present experiment.

Whilst the primary aim of Experiment 4 was to investigate whether contrast class information remained an important predictor of DG task success using the “three different numbers” rule, it was also useful to explore other accounts of the DG facilitation effect, as these have also not been tested in the literature using different DAX rules to the standard “ascending numbers” one. It should be noted, however, that the use of the “three different numbers” rule has a major impact on the way in which generated triples can be coded, which in turn means that predictions that derive from
different theories also differ markedly from predictions that are established in the context of Wason’s original rule. For example, triple heterogeneity theory (Vallée-Tourangeau et al., 1995) claims that success is associated with a wide exploration of the hypothesis space as indexed by the range of triples produced. While this prediction remains intact, the coding of triples used in Experiments 1 to 3 reflects the relational nature of the target rule and centres on positive triples (i.e., ascending triples of two types: posvars and constant positives) and negative triples (i.e., the eight types of non-ascending triples). The more general rule used in Experiment 4 requires a different coding system, with positive triples being any triple in which no number is repeated regardless of the relational properties of the triple, and negative triples having one number repeated at least once. What this means, however, is that the negtype variable has a much smaller range of examples contained within it (i.e., two numbers the same, or three equal numbers). As such it is clear that the negtype measure of heterogeneity becomes of limited use as an index of the breadth of search of the hypothesis space. The negative feedback measure suffers from the same limitation. However, given that the DAX seed triple (2-4-6) remains the same in Experiment 4 as in the standard DG task then it could be assumed that similar types of hypotheses to those for the standard rule would be entertained by participants (at least in the initial stages of hypothesis formation and testing). For this reason it was decided to employ the triptype measure introduced in Experiment 3 in the present experiment. While this measure is based on relational properties of triples generated it is not restricted to the production of triples which do not match the target hypothesis and therefore overcomes this limitation of the negtype measure.

With the variation to the rule and the coding of triples it is useful to make explicit the predictions made by competing theories. Previous experiments reported here have demonstrated that neither Evans’ (1989) positivity bias theory nor Wharton et al.’s (1993) goal complementarity theory provide a convincing account of the DG facilitation effect. However, to maintain control across experiments in terms of possible linguistic effects of feedback, any “three different number” triples produced by participants were designated as DAX triples, and those not fitting the rule were labelled MED in both the SG and DG conditions. Goal complementarity theory predicts that in the standard ascending rule a posvar triple allows the relaxation of the “increasing by twos” constraint. Such a triple would clearly not be expected to have any impact on success using the more general rule. Given that in the SG condition similar triple generation profiles to those in the standard “ascending numbers” rule would be expected it is unclear what particular type of triple would lead to relaxation of the participant’s rule as there appears to be no equivalent to a posvar using the three different numbers rule, thus testing of the goal complementarity account using triple
Experiment 4: Extending Contrast Class Theory to Different DAX Rules

analysis would be difficult. Experiment 3 also provided powerful evidence against goal complementarity theory; this account was therefore not tested in Experiment 4. As with Experiment 3 information quantity was controlled for by asking all participants to produce exactly 10 triples before announcing a rule.

The primary predictions of triple heterogeneity theory (Vallee-Tourangeau et al., 1995) remain intact: DG rules will elicit a wider exploration of the hypothesis space as indexed by a greater range of triples, and this increased range of triples will be associated with success on the task. Contrast class theory (Experiment 1 to 3) predicts that it is the production of a triple of the “psychological set complement” (Oaksford & Stenning, 1992) which would be predictive of success on the task; thus contrast class theory predicts that it is the production of a “three equal numbers” triple which would be the key determinant of success on the task.

In summary, then, Experiment 4 aimed to discriminate between the triple heterogeneity and contrast class accounts by exploring the triple generation patterns of those given SG and DG instructions and between those of solvers and non-solvers on Wason’s (1960) 2-4-6 task, using the more difficult “three different numbers” rule.

6.2 Method

6.2.1 Participants

Forty students at sixth-form colleges local to the University of Derby voluntarily took part in the study. None of the participants had received training in logic or reasoning.

6.2.2 Design

An independent measures design was employed, in which the instructions were manipulated such that half of the participants received SG instructions and the remainder DG instructions.

6.2.3 Materials and Procedure

Participants were tested in groups of up to four, with each participant producing written triples and receiving written feedback. Participants were provided with a sheet on which to record triples, feedback and their best guess at the rule after producing 10 triples. As with previous experiments participants were informed that they would be permitted only a single attempt at stating the rule.

Participants in the SG condition received instructions relating to a unique to-be-discovered rule: “I have in mind a rule that specifies how to make up sequences of
Chapter 6

Experiment 4: Extending Contrast Class Theory to Different DAX Rules

three numbers (triples), and your task is to discover this rule. Triples that fit my rule are called DAX triples and other triples are called MED triples. To start you off, I can tell you that 2–4–6 is a DAX triple. In order to discover my rule you should produce further number triples, and I will tell you whether they are DAX triples, or whether they are MED triples”. In the DG condition the instructions were varied to refer to two rules, called DAX and MED: “I have in mind a rule that specifies how to make up sequences of three numbers (triples), and your task is to discover these rules. Triples that fit one of my rules are called DAX triples and those that fit my other rule are called MED triples. To start you off, I can tell you that 2–4–6 is a DAX triple. In order to discover my rule you should produce further number triples, and I will tell you whether they are DAX triples, or whether they are MED triples”. Participants in both conditions were asked to test exactly ten triples after which they should write down their best guess at the rule(s). The DAX rule in both conditions was “any three different numbers” while MED triples were of any other form.

6.3 Results

6.3.1 Task Success across Condition

Before embarking on an analysis of the types of triples produced across conditions and solvers versus non-solvers it was essential to first establish that DG instructions do indeed have a facilitatory effect when the more general “three different numbers rule” is used. The data presented in Table 6.1 show the success rates of participants in the DG and SG conditions. It is clear that participants in the SG condition did indeed find the task very difficult with only one participant correctly announcing the rule. It is also apparent that there was there is a strong facilitatory effect, with 60% solving the task in the DG condition, compared to 5% in the SG condition; this difference was found to be highly reliable, $\chi^2 (1, N = 40) = 13.79, p < .001, w = .58$. 

- 93 -
# Table 6-1

**Frequency of Solvers and Non-Solvers across Conditions, Broken Down According to the Presence versus Absence of at Least One Contrast Class Triple**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Presence of Contrast Class Triple</th>
<th>Solvers</th>
<th>Non-Solvers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Present</td>
<td>Absent</td>
<td>Total</td>
</tr>
<tr>
<td>SG</td>
<td>1</td>
<td>0</td>
<td>1 (5%)</td>
</tr>
<tr>
<td>DG</td>
<td>10</td>
<td>2</td>
<td>12 (60%)</td>
</tr>
</tbody>
</table>

### 6.3.2 Triple Generation Profile of Successful versus Non-Successful Participants

Having demonstrated the facilitatory effect of DG instructions it was apparent that the dataset was a suitable vehicle for exploring theories of DG facilitation. Recall that contrast class theory predicts that task success should be associated with the generation of a triple of the "three equal numbers" form. To explore this prediction data were collapsed across condition and success crossed with the presence or absence of a contrast class triple (see Table 6.1). A chi-square test of independence showed that the association between the production of a three equal numbers triple and successful rule discovery was indeed the case, $\chi^2 (1, N = 40) = 27.36, p < .001, w = .83$, with only two participants solving the task in the absence of such a triple. It was also important to check that DG instructions promoted the generation of a contrast class triple. Again a strong effect was found, $\chi^2 (1, N = 40) = 11.91, p = .001, w = .55$, with 55% of those in the DG condition producing such a triple, compared to 5% in the SG condition (Table 6.1). Thus the data presented here again offer support for the contrast class account of DG facilitation.

To examine the prediction that only contrast class triples promote task success, a chi-square test of independence was conducted on the dataset depicted in Table 6.2, which crossed instructions with production of triples that were neither all different or all the same (i.e., triples with two equal numbers). This showed that instruction type was not associated with the production of these types of triples, Fisher’s Exact = .661 (low expected frequencies means that assumptions of a chi-square were violated). Further analysis in which success was collapsed across instructions showed that these
types of triples were not associated with success on the task, Fisher's Exact = .075.

Taken together these findings attest to the importance of contrast class triple production in mediating task success.

Table 6-2

*Frequency of Solvers and Non-Solvers across Conditions, Broken Down According to the Presence versus Absence of at Least One Triple Containing 2 Equal Numbers*

| Presence of Two Equal Numbers Triple | Solvers | | | | Non-Solvers | | |
| Condition | Present | Absent | Total | Present | Absent | Total |
| SG | 1 | 0 | 1 (5%) | 2 | 17 | 19 (95%) |
| DG | 4 | 8 | 12 (60%) | 0 | 8 | 8 (40%) |

The contrast class account of DG facilitation focuses on the production of a single triple of the "three numbers the same" form. Triple heterogeneity theory (Valleé-Tourangeau et al., 1995) on the other hand focuses on the range of triples produced by participants. As discussed in the introduction it is difficult to code triples meaningfully in relation to the target rule used in this experiment since the non-relational form of the target hypothesis makes the relational emphasis of the triple-coding scheme used thus far somewhat restrictive. However, given the identical seed triple it seems reasonable to assume that similar hypotheses will be considered by participants seeking the target rule used in this experiment as those seeking the standard "ascending numbers" rule, at least in the initial stages of their hypothesis testing. Also as noted in the introduction the definition of a negative triple is different in this experiment compared with Experiments 1 to 3. In the current study only two types of negative triple are possible: a triple with two equal numbers and a triple with three equal numbers. Thus it was argued that the range of negtypes in this experiment is too limited to provide a sound test of triple heterogeneity theory whilst the negative feedback measure suffers from a similar limitation. By contrast, the triptype index introduced in Experiment 3 does provide a reasonable measure of the range of triples being tested, therefore this measure of triple heterogeneity was used in the present analyses. In addition the constant positive measure is also analysed as inspection of the rules announced by unsuccessful solvers showed a high rate of "numbers ascending by two" types, thus it could be argued that the constant positive measure remains an index of limited hypothesis testing.
Table 6.3 shows the measures of triple heterogeneity across success. There were significant effects of both measures across success with solvers producing fewer constant positive triples than non-solvers, $t(38) = 4.29, p < .001, d = 1.2$, but more triptypes, $t(38) = 5.85, p < .001, d = 1.98$. These results offer support for the triple heterogeneity theory of DG facilitation.

<table>
<thead>
<tr>
<th>Measure of Heterogeneity</th>
<th>Non-Solvers</th>
<th>Solvers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant Positive Triples</td>
<td>2.92 .50</td>
<td>6.96 .61</td>
</tr>
<tr>
<td>Triptype Score</td>
<td>5.08 .40</td>
<td>2.40 .25</td>
</tr>
</tbody>
</table>

### 6.3.3 Triple Types across Condition

While the previous analyses showed that solvers produced a wider range of triples there remains the need to ascertain that the generation of these triples was promoted by DG instructions. A second set of analyses was therefore performed in which the effect of instruction type on triple generation profiles was examined. These analyses showed an effect of instruction type on triple generation profiles with those in the SG condition producing more constant positive triples, $t(38) = 4.32, p < .001, d = 1.13$, but fewer triptypes, $t(38) = 4.86, p < .001, d = 1.66$, demonstrating that DG instructions promoted the generation of a wide range of triples as suggested by triple heterogeneity theory of DG facilitation (see Table 6.4).
Table 6-4

<table>
<thead>
<tr>
<th>Instructions</th>
<th>SG</th>
<th>DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure of Heterogeneity</td>
<td>Mean</td>
<td>SE</td>
</tr>
<tr>
<td>Constant Positive Triples</td>
<td>7.55</td>
<td>0.74</td>
</tr>
<tr>
<td>Triptype Score</td>
<td>2.15</td>
<td>0.33</td>
</tr>
</tbody>
</table>

6.3.4 Discriminating between the Triple Heterogeneity and Contrast Class Accounts

The triple heterogeneity theory emphasises that production of a wide range of triples is instrumental in facilitating success, whereas the contrast class theory dismisses the importance of variety of triple types, instead proposing that production of at least one descending triple is the critical determinant for rule discovery. To clarify the relative roles of triptypes and the presence versus absence of a contrast class triple for task success the dataset was modelled using logistic regression. An initial analysis in which the dichotomous predictor variable of contrast class triple present versus absent was regressed onto solution success was highly significant, $B = 4.96$, Wald = 15.16, $p < .001$. Indeed, participants generating a three equal numbers triple were 142 times more likely to discover the DAX rule than those who didn’t generate such a triple. A second model in which triptype score was regressed onto solution success also reached significance, $B = 1.83$, Wald = 7.30, $p = .007$. A final model in which the presence versus absence of a contrast class triple and triptype score was regressed onto success was run. This showed that when both variables were regressed onto success the presence of a contrast class triple remained predictive of success, $B = 6.19$, Wald = 5.27, $p = .02$, whereas the triptype measure no longer achieved significance $B = 2.37$, Wald = 3.72, $p > .05$. These analyses show that while both contrast class triples and triple variety (as indexed by the triptype measure) have a role to play in solving the task, for the present dataset it appears that the production of a contrast class triple is the more critical determinant of successful rule discovery.
6.4 Discussion

This experiment was designed to test the generalisability of contrast class theory to DAX rules other than the standard ascending numbers rule. Analysis of the data has shown that DG instructions are effective in facilitating performance on this more difficult version of the task, an important precursor to exploring theoretical explanations of a general contrast class account. In addition, analysis of the triple generation profiles of successful and non-successful solvers demonstrated that production of a contrast class triple is closely associated with success on the task. It is also clear that DG instructions promoted the production of contrast class triples. Modelling the data using logistic regression showed that instruction type (SG vs. DG) was predictive of the production of such a triple and also that generation of a contrast class triple was predictive of success, thus offering further support for the contrast class account.

An important finding evident here is that the generation of a “two equal numbers” triple was not associated with success. This finding points to the importance of the psychological set complement (Oaksford & Stenning, 1992) aspect of contrast class theory. The dataset presented here shows that DG instructions induced the generation of triples which were examples of psychological set complements rather than logical opposites (i.e., “three equal numbers” rather than “two or more equal numbers”), and that these contrast class triples were associated with success, whereas the apparently equally informative “two equal numbers triples” were not. However there were few participants who produced these latter sorts of triples, so it is difficult to draw conclusions from the data with regard to this.

The second theory of interest was that of triple heterogeneity. Initial data analysis offered some support for this theory as it was evident both that DG instructions promoted a wide range of triples, and that this wide range of triples was associated with success on the task. Modelling of the data using logistic regression showed that instruction type (SG versus DG) was predictive of the generation of a wide range of triples and that producing a wide range of triples was predictive of success. However this effect disappeared when production of a contrast class triple was controlled for. Results from this study therefore offer some support for triple heterogeneity theory although contrast class theory appears to provide a better account of the data.

To summarise, the data presented here offer further support for the contrast class account of DG across a different DAX rule, which represents an important generalisation of the theory. Further, there was some support (albeit slightly weaker) for triple heterogeneity theory. One weakness of investigations of both of the accounts
Experiment 4: Extending Contrast Class Theory to Different DAX Rules

discussed here is that testing of the theories depends on a *post hoc* analysis of the triple generation profiles of participants rather than direct manipulations of the types of triples used in hypothesis formation and testing. A stronger test of the theories could be made by manipulating the types of information available to participants, an approach which is adopted in the following experiment.
7.1 Introduction

The experiments that have been reported in this thesis so far have pointed to the central role of descending triples in successful rule discovery in the 2-4-6 task. The observation that solvers in both the SG and DG paradigms are more likely to have generated at least one descending triple has, in turn, led to the development of a contrast class account of task success, whereby the production of a descending triple provides a salient contrast class cue to the ascending nature of the DAX rule. It has also been proposed that DG instructions are more likely to induce participants to produce a descending triple, thereby providing a novel theoretical explanation of the general DG facilitation effect.

The contrast class theory of DG facilitation is based on *post hoc* analyses of hypothesis testing strategies and, as such, it is not possible to make causal claims as to the importance of contrast class information. The primary aim of the present experiment was, therefore, to explore further the contrast class theory by manipulating the information available to participants in order to test the theory more directly. To this end, in addition to the standard 2-4-6 DAX exemplar participants were also provided with a MED exemplar that provided either a “useful” or a “non-useful” contrast class cue. Half of the participants were therefore told that an example of the MED rule was the triple 6-4-2 (useful contrast class information). It was predicted that this MED exemplar would promote identification of the DAX rule since 6-4-2 and 2-4-6 are oppositional on the salient – and crucially relevant – dimension of “ascending” versus “descending”. The other half of the participants were told that an example of the MED rule was the triple 4-4-4 (non-useful contrast class information) with the prediction that this MED triple would not promote successful DAX discovery as 4-4-4 and 2-4-6 are oppositional on the salient – but non-relevant – dimension of “three identical numbers” versus “three different numbers”.

In addition to the introduction of the notion of contrast class information playing a crucial role in successful rule discovery evidence presented within this thesis has also offered support to the triple heterogeneity account of DG facilitation (Vallée-Tourangeau et al., 1995). In Experiment 3, for example, it was shown that successful solvers produced a greater range of triples than non-solvers, and that those given DG instructions reliably produced a wider range of triples. In the present experiment, then, *post hoc* analyses of triple generation profiles were likewise planned as a means to explore the triple heterogeneity account. It was
## Table 7-1

### Theories of Dual Goal Facilitation, Including Predictions and Empirical Tests

<table>
<thead>
<tr>
<th>Account</th>
<th>Features</th>
<th>To Test</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrast class cue theory</td>
<td>DG instructions promote production of descending triples</td>
<td>Compare presence of descending triples across instruction mode and for solvers versus non-solvers</td>
<td>Production of a descending triple mediates between instruction mode and success on the task</td>
</tr>
<tr>
<td>(Experiment 1)</td>
<td>Descending triple generation is associated with task success</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triple heterogeneity theory</td>
<td>DG facilitation depends upon the production of a wide variety of triples types</td>
<td>Analyse the variety of triple types across conditions as well as for solvers versus non-solvers</td>
<td>Variety of triples produced mediates between type of instruction and success on the task</td>
</tr>
<tr>
<td>(Vallée-Tourangeau, Austin, &amp; Rankin, 1995)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal complementarity theory</td>
<td>Only complementary goals cause DG facilitation.</td>
<td>Manipulate the complementarity of the DAX and MED rules</td>
<td>DG facilitation is not evident when non-complementary rules are used</td>
</tr>
<tr>
<td>(Wharton et al., 1993)</td>
<td>Task success is dependent on the production of at least one posvar triple</td>
<td>Compare presence of posvars across conditions as well as for solvers versus non-solvers</td>
<td>Production of a posvar triple mediates between type of instruction and success on the task</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
also decided to err on the side of caution and perform post hoc analyses of posvar
generation as an additional test of the goal complementarity theory. Additionally, as
with Experiment 3 and the associated follow-up study presented in the same chapter,
an instruction that exactly 10 triples were to be produced was used in this experiment
so as to control for any effect that information quantity might have on task success.

7.2 Method

7.2.1 Participants

Forty-three first-year psychology students from the University of Derby took part
in the study on a voluntary basis. None had received any teaching relating to reasoning
or logic.

7.2.2 Design

An independent-measures design was employed with the manipulation reflecting
the usefulness of the contrast class cue (CCC) that was salient in the presented
example of the MED rule. One group of participants received a useful CCC (6-4-2)
and the other group were given a non-useful CCC (4-4-4). Participants were randomly
assigned to the two conditions. The DG paradigm was used in both conditions of this
experiment.

7.2.3 Procedure

Participants were tested individually in a quiet laboratory. Standardised DG
instructions were read out to all participants as follows: “I have in mind two rules that
specify how to make up sequences of three numbers (triples), and your task is to
discover these rules. Triples that fit one of my rules are called DAX triples and those
that fit my other rule are called MED triples...”. All participants were given 2-4-6 as
an example DAX triple. Those in the useful CCC condition were given 6-4-2 as a
MED exemplar, while those in the non-useful CCC condition were given 4-4-4 as a
MED exemplar. Participants were then asked to produce exactly 10 triples, and they
received feedback for each triple in the form of “DAX” or “MED”. After 10 triples
had been generated participants were asked to write down their best guess at the two
rules.
7.3 Results

7.3.1 Solution Success across Conditions

Table 7.2 shows the frequency of correct rule announcements for the DAX rule across experimental conditions. It is clear that the usefulness of the CCC that had been provided had a dramatic effect on success rates for this task, with success rates being over three times higher for those participants given useful contrast class information. A chi-square analysis showed this effect to be highly significant, \( \chi^2 (1) = 12.44, N = 43, p < .001, w = .54 \). These success rates are very similar to those typically reported in the literature for the DG and SG paradigms respectively (e.g., Wharton et al., 1993). This observation suggests that it may not be DG instructions per se that lead to task success in the DG paradigm, but rather that DG instructions facilitate participants' production of a salient contrast class that, in turn, promotes successful discovery of the target DAX rule.

Table 7-2

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Solver</th>
<th>Non-solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Useful CCC</td>
<td>23</td>
<td>74</td>
<td>26</td>
</tr>
<tr>
<td>Non-Useful CCC</td>
<td>20</td>
<td>20</td>
<td>80</td>
</tr>
</tbody>
</table>

7.3.2 Production of Descending Triples

The purpose of this study was to test the idea that the key to success on the DG 2-4-6 task relates to the availability or discovery of useful contrast class information that facilitates identification of the potential scope of the DAX rule. In Experiments 1 to 4, it was noted that successful solvers are those who uncover at least one descending triple during their hypothesis testing, and that DG instructions promote the production of at least one such descending triple. It was therefore decided to examine the present dataset for any effect of the presence versus absence of the participants' production of a descending triple on their task success. Table 7.3 presents data collapsed across the useful versus non-useful CCC manipulation. This shows that while the majority of participants produced at least one descending triple, there was no single instance of a participant solving the task in the absence of a descending triple.
A chi-square analysis indicated that this effect was highly reliable, $\chi^2 (1) = 12.44$, $N = 43, p < .001, \omega = .54$.

Table 7-3

**Percentage of Correct DAX Announcements by Presence versus Absence of at Least One Descending Triple**

<table>
<thead>
<tr>
<th>Descending Triple</th>
<th>N</th>
<th>Solver</th>
<th>Non-solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>10</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Present</td>
<td>33</td>
<td>64</td>
<td>36</td>
</tr>
</tbody>
</table>

The data were further explored to ensure that the type of CCC that had been provided had a reliable influence on whether or not at least one descending triple was produced (see Table 7.4). This analysis (collapsing across solver vs. non-solver) showed that all participants receiving a useful CCC produced at least one descending triple, while only 50% of participants receiving a non-useful CCC produced a descending triple, $\chi^2 (1) = 14.99$, $N = 43, p < .001 \omega = .59$. It can thus be concluded that the CCC manipulation was successful in terms of its capacity to induce the generation of descending triples.

Table 7-4

**Percentage of Participants Producing at Least One Descending Triple by Condition**

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Absent</th>
<th>Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>Useful CCC</td>
<td>23</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Non-Useful CCC</td>
<td>20</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

7.3.3 Analysis of Triple Types Produced across Condition

One stated aim of the present study was to further explore other accounts of DG facilitation such as Vallée-Tourangeau et al.’s (1995) triple heterogeneity theory. According to this theory task success is dependent on the production of a wide range of triples, with this wide range of triples being promoted by the use of DG instruction.
Predictions derived from the theory are somewhat problematic in the present study; all participants were given DG instructions, which would lead to the prediction that there would be no difference in the number of triples produced across condition. On the other hand, the differential performance across conditions would tend to suggest that those in the more successful “useful” contrast class condition would produce a wider variety of triples.

Table 7-5

Participants' Triple-Generation Profile across Conditions

<table>
<thead>
<tr>
<th>Measure of Heterogeneity</th>
<th>Type of Contrast Class Cue</th>
<th>Mean</th>
<th>SE</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant positives</td>
<td>Useful</td>
<td>4.78</td>
<td>.26</td>
<td>3.70</td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>Non-useful</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posvars</td>
<td></td>
<td>0.48</td>
<td>.20</td>
<td>0.90</td>
<td>.28</td>
</tr>
<tr>
<td>Number of triples receiving negative feedback</td>
<td></td>
<td>4.61</td>
<td>.20</td>
<td>5.25</td>
<td>.40</td>
</tr>
<tr>
<td>Negtypes</td>
<td></td>
<td>1.26</td>
<td>.13</td>
<td>2.80</td>
<td>.46</td>
</tr>
<tr>
<td>Triptypes</td>
<td></td>
<td>2.52</td>
<td>.16</td>
<td>4.15</td>
<td>.42</td>
</tr>
<tr>
<td>Descending triples</td>
<td></td>
<td>4.04</td>
<td>.26</td>
<td>0.85</td>
<td>.23</td>
</tr>
</tbody>
</table>

To explore these ideas a series of t-tests were pursued to examine aspects of participants’ triple generation profiles across conditions (Table 7.5). Reliable differences were found between conditions in the number of constant positive triples, \( t(41) = 2.12, p = .04, d = .62 \), the number of negtypes, \( t(41) = 3.41, p = .001, d = .93 \), the number of triptypes \( t(41) = 3.76, p = .001, d = 1.00 \), and the number of descending triples, \( t(41) = 9.14, p < .001, d = 1.62 \). The descending triple measure is, of course, not a measure of heterogeneity, and as such is of little relevance to triple heterogeneity theorists; it is, however, included here since the generation of such triples is predicted by the contrast class theory. The mean triple variety scores for participants were lower in the useful contrast class group than the non-useful contrast class group on all measures except for constant positives, where the mean was higher. The differences in the types of triples produced across conditions tend to militate against the triple.
heterogeneity theory, which would suggest that since all participants were given DG instructions then triple generation profiles would be similar across conditions.

7.3.4 Analysis of Triple Types Produced across Success

To examine the triple heterogeneity theory more closely a second set of analyses explored the differences in the triple generation profiles of successful versus non-successful participants (see Table 7.6). These showed that there were no differences evident across solvers and non-solvers on any measure of triple heterogeneity, with the exception of descending triples, \( t(41) = 5.18, p < .001, d = 1.25 \).

Table 7-6

<table>
<thead>
<tr>
<th>Measure of Heterogeneity</th>
<th>Success on Task</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non-Solver</td>
<td>Solver</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
</tr>
<tr>
<td>Constant positives</td>
<td>3.86</td>
<td>.49</td>
<td>4.71</td>
</tr>
<tr>
<td>Posvars</td>
<td>0.64</td>
<td>.25</td>
<td>0.71</td>
</tr>
<tr>
<td>Number of triples receiving negative feedback</td>
<td>5.18</td>
<td>.37</td>
<td>4.62</td>
</tr>
<tr>
<td>Negtypes</td>
<td>2.32</td>
<td>.44</td>
<td>1.62</td>
</tr>
<tr>
<td>Triptypes</td>
<td>3.55</td>
<td>.41</td>
<td>3.00</td>
</tr>
<tr>
<td>Descending triples</td>
<td>1.36</td>
<td>.37</td>
<td>3.81</td>
</tr>
</tbody>
</table>

7.4 Discussion

The primary aim of Experiment 5 was to test directly the contrast class account of DG facilitation on the 2-4-6. The contrast class account posits that it is not DG instructions per se which promote task success, but rather that DG instructions promote the production of a salient contrasting descending triple, and it is the production of such a triple which facilitates performance. It was, therefore, predicted that the DG facilitatory effect would be suppressed by giving participants non-relevant information about the MED rule which would inhibit the production of a salient
contrast class triple. Results of the study do indeed support this notion, with success rates of participants given the non-useful contrast class information approximately equivalent to those demonstrated in the standard, SG form of the task. This is only the second time that non-facilitated performance has been reported using DG instructions (cf. Wharton et al., 1993). Further support for the contrast class theory was gained from analyses of the triples produced by participants. These showed that there was no single case of solving the task in the absence of a descending triple, and that all participants in the useful contrast class group did indeed produce at least one descending triple, whereas only half of the participants given the non-useful contrast class cue did so, thus offering support to the notion that it is the production of a descending triple which is crucial to success on the task.

A secondary aim of the study was to explore other theories of DG facilitation. By requiring all participants to generate exactly 10 triples the experiment controlled for information quantity (Wharton et al., 1993) whilst analysis of the 10 triples produced enabled the quantification of key aspects of the triple heterogeneity account (Vallée-Tourangeau et al., 1995). The results showed differential performance on the task despite controlling for the number of triples produced, whilst analyses of the triples indicated that contrary to the prediction of the triple heterogeneity account solvers varied from non-solvers on only one aspect of the triple production, that is, the number of descending triples produced. As such, solvers and non-solvers generated a similar number and range of triple types, thus indicating that DG facilitation cannot be attributed to a creative exploration of the problem space.

The analyses of triple generation profiles presented here seem to indicate that participants appreciate the value of a descending triple in delimiting the scope of the DAX rule, and that when given an example of such a triple they engage in less exploration of the problem space than those given the non-useful contrast class triple. These findings tend to indicate that it is not a wide exploration of the hypothesis space which leads to success, but rather that it is this wide search which mediates between DG instructions and the production of a useful contrast class triple and therefore success. It is possible that rather than being competing accounts of DG facilitation, the contrast class theory and the triple heterogeneity theory may both represent important aspects of people’s hypothesis testing strategies (cf. Roberts, 2000; Siegler & Chen, 2002)

The data presented here appear to show a convincing case for the contrast class account of DG facilitation. What remains, however, is the need to provide the details of a fully-fledged theoretical framework to underpin the contrast class account. In Chapter 5 it was argued that the contrast class account could be accommodated within
what appears to be the most psychologically plausible account of behaviour on the SG version of the 2-4-6 task, that is, Oaksford and Chater's (1994) Iterative Counterfactual Model (ICM), itself a development of Farris and Revlin's (1989ab; 1989ba) earlier counterfactual strategy. The ICM focuses on how hypotheses are created rather than on how they are tested, with one key aspect being how hypotheses are revised when falsifying evidence is obtained. The ICM is a normative model and thus describes the strategy used by successful hypothesis testers and applies in its original format to the SG paradigm. How then does the model explain the facilitatory effect of DG instructions, in particular the contrast class account of the effect?

To answer this question an extension to the ICM is proposed (see Figure 7.1). In this revised model the processes by which the DAX hypothesis is generated and tested remain in place, and it is only when the participant has tested and become reasonably confident in their DAX hypothesis (Box 14) that the extension to the model becomes activated: after having produced their DAX hypothesis the participant moves on to discovering the MED rule. Unlike the initial stages of DAX hypothesis discovery they do not have a seed triple and they must therefore generate the MED hypothesis using a different strategy. They do know that DAX and MED are different and that DAX seems to be “numbers ascending by 2”’. It seems reasonable to assume that participants will use this information to generate their initial MED hypothesis so that it differs from their DAX hypothesis as much as possible (i.e., it is the “opposite” of DAX). Note that this hypothesis contains two axes, “ascending” and “by 2”. Iterative counterfactual testing is concerned with exploring one common feature at a time; it is, therefore, reasonable to suppose that when generating their MED hypothesis participants will focus on one axis of the DAX hypothesis and vary that. Evidence for the view that a single feature is considered at a time comes from Tschigi (1980) who showed that when testing hypotheses reasoners used a strategy of varying only one aspect of their materials at a time. The participant therefore selects one of these properties (e.g., “numbers rising by 2”, Box 15), generates a contrast of this property (“numbers rising by 3”, Box 16) and produces an instance (I) of their MED hypothesis (Box 17), which is tested (Box 18). Feedback for this (I) is, of course, DAX. Their tentative MED H must therefore be rejected and I added to the list of the DAX instances (Box 22). The participant is now aware that their DAX hypothesis is wrong and moves back to discovering the DAX rule using iterative counterfactual testing (Box 2). Once satisfied with their amended DAX hypothesis (Box 14) they then once again move to the generation and testing of the MED hypothesis. Conversely the participant may vary the “ascending” aspect of their DAX hypothesis and generate a “decreasing by 2” MED hypothesis, subsequently producing an instance of their MED hypothesis (e.g., 6-4-2, Box 17). On testing (Box 18) they will receive MED feedback.
Chapter 7

Experiment 5: Testing Contrast Class Theory

The participant may now move into a MED positive sub-loop, generating and testing instances of their MED hypothesis until (Box 19) they have reached the criterion level set for sufficient confidence in their MED hypothesis (Box 21). The process is then terminated and their MED hypothesis is compared to the DAX hypothesis (Box 22) for any common properties (Box 23). In the example given the common property of "by 2" is noted as being common to both hypotheses, leading to the generation of a complement to that property (Box 3) and an iteration of the DAX and MED hypothesis testing procedure as outlined above. If no common properties of the DAX and MED hypotheses are perceived then the rules are announced. It is this process of counterfactual testing of the DAX hypothesis and production of contrasting MED hypotheses which leads to the facilitatory effect of DG instructions.

Supporting evidence for an extended ICM that incorporates ideas of contrast class identification is provided by the solution success results of Experiment 5. In relation to the non-useful CCC condition, for example, it had been expected that participants who were given the illustrative 4-4-4 MED triple would generate a “three equal numbers hypothesis” for MED, and would, therefore, be lured toward considering a (non-relevant) “three different numbers” hypothesis for DAX. The results did indeed show instances of participants going as far as announcing a final DAX rule as being “three different numbers” in the non-useful MED condition. This supports the notion that reasoners are highly susceptible to focusing on the apparent relevance of available contrast class cues during their rule induction and hypothesis testing. It is envisaged that the provision of non-useful contrast class information hindered performance on the task by interfering with the MED hypothesis stage in which a single property of the DAX hypothesis is varied, thus the salient useful contrasting MED hypotheses are not generated. Furthermore, the instructions in use did not explicitly state that the two rules were exclusive and exhaustive, so it may be that reasoners were pursuing two SG rules rather than two related rules. This second explanation is somewhat undermined, however, by noting that those participants receiving useful contrast class information were given exactly the same instruction but successfully solved the task and therefore did not appear to be looking for two unrelated rules.

The data from this study demonstrate the importance of contrast class information in solving the rule discovery task. However, as noted earlier, the MED exemplar provided to participants in the useful contrast class cue condition was oppositional to the DAX rule on a single dimension only, that of ascending versus descending. The universe of descending triples is, however, infinite and contains
Figure 7-1

Extended Iterative Counterfactual Model

Start

Add seed triple to list L of instances

2. Select a common property of L as H

3. Generate Complement of H

4. Generate instance H

5. Is f an instance of H?
   YES
   Generate Instance (i) of H
   NO
   Select Property of DAX H

6. Is f an instance of T?
   YES
   Generate contrast of property (MED H)
   NO
   Compare DAX H and MED H

7. Reject H and H' and Add f to L

8. Add i to L

9. Any perceivable common properties?
   YES
   No MED testing
   NO
   Rejected MED H and add i to list

10. Is i an instance of MED H?
    YES
    MED testing
    NO
    MED positive sub loop

11. Add i to L

12. Reject H and do not add i to L

13. Select Property of DAX H

14. DAX H

15. MED H

16. Generate Instance (i) of MED H

17. Add i to L

18. Is i an instance of MED H?
    YES
    MED testing
    NO
    MED positive sub loop

19. Is f an instance of T?
    YES
    MED testing
    NO
    MED positive sub loop

20. Any perceivable common properties?
    YES
    No MED testing
    NO
    Rejected MED H and add i to list

21. Merging DAX H and MED H

22. MED H

23. QUIT

Chapter 7
Experiment 5: Testing Contrast Class Theory
triples with a wide variety of properties, for example, only odd numbers, unequal interval between integers, prime numbers etc. It is also notable that the useful contrast class exemplar was not the logical opposite of the most commonly reported DAX hypothesis (i.e., “numbers rising by 2”), but rather represents its psychological set complement (i.e., “numbers descending by 2”). Oaksford (2002) has argued that a contrast class is “made up of the most likely or relevant members of the complement set” (p.140), and is thus not necessarily a logical concept. It could be argued that a descending triple which varied from the original DAX exemplar on a multiplicity of properties would not constitute a psychological contrast class, and would therefore not facilitate performance on the task. This argument is in line with an implicit prediction of the extended ICM: that should MED information which varies on a variety of axes be provided then this should interfere with the working of the model, thereby hindering performance on the task. It was, therefore, decided to test this prediction by manipulating the MED information such that whilst it differed from DAX information on the salient ascending/descending axis, it also varied on other axes.
Chapter 8
Experiment 6: Manipulating Contrast Class Information across More than One Axis

8.1 Introduction

As noted in Chapter 7, the extended ICM model presented in this thesis suggests that when given DG instructions the participant first generates and tests their DAX hypothesis to a preset level of confidence, and then turns to generating and testing the MED hypothesis. It has also been proposed that the MED hypothesis is typically drawn from two pieces of information: the reasoner’s DAX hypothesis and their knowledge that the two rules are different. As such, it is assumed that participants generate a MED hypothesis which is “opposite” to their DAX hypothesis. The model asserts that this latter process is achieved by the participant selecting a single property of their DAX hypothesis and then generating a contrast to that property.

One implicit assumption of the proposed model is, therefore, that should a MED exemplar be given which varies on more than one axis then this would interfere with processing and hinder performance on the DG task. To this end it was decided to manipulate the MED information given such that 9-8-2 was presented as an illustrative MED triple as this contrasts with 2-4-6 on: (1) several, non-relevant dimensions (e.g., “mixed odd and even numbers” versus “only even numbers”, “unequal intervals” versus “equal intervals”, and “middle number is not arithmetic mean of the outer numbers” versus “middle number is the arithmetic mean of the outer numbers”); and (2) a single, relevant dimension (i.e., “descending numbers” versus “ascending numbers”). In addition, given that only one previous instance of DG instructions not facilitating performance has been reported (Wharton et al., 1993), it was decided to collect a further set of data from participants given 4-4-4 as non-useful CCC information, to check the replicability of performance levels observed in Experiment 5. Finally, a control group was given 6-4-2 as a MED exemplar. It was predicted that the multiplicity of the available contrast class cues in the 9-8-2 condition would hinder performance on the DG version of the task as compared to the useful contrast class condition.
8.2 Method

A further 64 first year students at the University of Derby took part in this study. The procedure employed was identical to that in Experiment 5 with the exception that participants were randomly assigned to one of three groups: one group was told that 6-4-2 was an example of a MED triple, the second group that 4-4-4 was a MED triple, and the final group was informed that 9-8-2 conformed to the MED rule.

8.3 Results and discussion

8.3.1 Success Rates across Conditions

Comparison of success rates across conditions shows that success in the “useful” CCC condition was double that achieved in the “multiple-axes” CCC condition, and three times higher than for the group receiving the “non-useful” CCC (see Table 8.1). A chi-square test of independence indicated that these differences were reliable, $\chi^2(2, \ N = 64) = 12.38, \ p = .002, \ w = .44$. Multiple chi-square analyses were performed with Bonferroni corrections applied (giving an alpha of .017) in order to determine where this difference lay. Results of these analyses showed there to be significant differences in success rates between those given the 6-4-2 exemplar and either the 4-4-4 exemplar, $\chi^2(1) = 11.48, \ N = 42, \ p = .001, \ w = .52$, or the 9-8-2 exemplar, $\chi^2(1) = 6.31, \ N = 42, \ p = .012, \ w = .39$. No difference was found between the 4-4-4 and 9-8-2 conditions, $\chi^2(1) = .98, \ N = 44, \ p = .32, \ w = .15$, (see Table 8.1).

Table 8-1

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Solver</th>
<th>Non-solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Useful CCC</td>
<td>20</td>
<td>75</td>
<td>25</td>
</tr>
<tr>
<td>Multiple Axes CCC</td>
<td>22</td>
<td>36</td>
<td>64</td>
</tr>
<tr>
<td>Non-Useful CCC</td>
<td>22</td>
<td>23</td>
<td>77</td>
</tr>
</tbody>
</table>

8.3.2 Production of Descending Triples

The focus of this experiment was to test the idea that a key feature of the extended ICM was that to be maximally useful to the reasoner the initial MED
Experiment 6: Manipulating Contrast Class Information across Axes

exemplar should vary only on a single axis that optimised the salience of descending triples. It was decided, therefore, that it would be prudent to assess the importance of the production of a descending triple for task success. Table 8.2 presents data collapsed across conditions. This table shows that while the majority of participants produced at least one descending triple, there was no single instance of a participant solving the task in the absence of a descending triple. A chi-square analysis indicated that this effect was highly reliable, \( \chi^2(1, N = 64) = 9.22, p < .01, \omega = .38 \).

Table 8-2

<table>
<thead>
<tr>
<th>Success</th>
<th>N</th>
<th>Solver</th>
<th>Non-solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descending Triple</td>
<td>Absent</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Present</td>
<td>54</td>
<td>28</td>
</tr>
</tbody>
</table>

8.3.3 Analysis of Triple Types across Condition

In order to explore further the strategies pursued across condition the data were analysed in terms of the types of triples produced (see Table 8.3). Recall that triple heterogeneity in its original form suggested that DG instructions promote success on the 2-4-6 task by inducing a wide search of the hypothesis space. However, analyses from Experiment 5 tended to suggest that a wide search of the hypothesis space mediated between DG instructions and the production of a contrast class triple, and that once such a triple had been identified the search of the hypothesis space was abandoned in favour of testing of the DAX and MED hypotheses. To discriminate between these ideas a series of one-way ANOVAs was undertaken which examined the triple generation profile of participants across condition. These ANOVAs showed that across conditions there were reliable differences in both the number of negtypes produced, \( F(2,61) = 4.11, p = .02, \eta^2 = .12 \), and the number of descending triples produced, \( F(2,61) = 15.36, p < .001, \eta^2 = .33 \). Post hoc Bonferroni tests showed that for the negtype measure this difference was between the useful CCC and the non-useful CCC conditions (\( p = .02 \)), while for descending triples there were differences between the useful CCC and the non-useful CCC conditions (\( p < .001 \)) and between the multiple axis CCC and the non-useful CCC conditions (\( p = .008 \)). These analyses indicate that when given a useful CCC triple, participants were able to recognise its
value and rather than searching the hypothesis space focussed on testing potentially informative descending triples.

Table 8-3

<table>
<thead>
<tr>
<th>Type of CCC Information</th>
<th>Measure of Heterogeneity</th>
<th>Mean</th>
<th>SE</th>
<th>Mean</th>
<th>SE</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constant positives</td>
<td>4.30</td>
<td>0.40</td>
<td>3.68</td>
<td>0.31</td>
<td>3.59</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Posvars</td>
<td>0.85</td>
<td>0.24</td>
<td>0.86</td>
<td>0.23</td>
<td>1.23</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Number of triples</td>
<td>4.90</td>
<td>0.34</td>
<td>5.55</td>
<td>0.29</td>
<td>5.00</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>receiving negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>feedback</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Negtypes</td>
<td>1.75</td>
<td>0.28</td>
<td>2.95</td>
<td>0.35</td>
<td>1.19</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Triptypes</td>
<td>3.25</td>
<td>0.34</td>
<td>4.41</td>
<td>0.39</td>
<td>4.00</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Descending triples</td>
<td>3.55</td>
<td>0.29</td>
<td>1.18</td>
<td>0.30</td>
<td>2.50</td>
<td>0.31</td>
</tr>
</tbody>
</table>

8.3.4 Analysis of Triple Types across Success

A further series of analyses explored the triple generation profiles of successful versus non-successful participants. Triple heterogeneity theory suggests that solvers will engage in a wider search of the hypothesis space than non-solvers, while contrast class theory is concerned with triples that reflect a reasoner’s “psychological set complement”. The triple generation profiles of solvers and non-solvers are shown in Table 8.4. This table shows that the only measure in which there was a reliable difference across conditions related to the number of descending triples produced, \( t(62) = 2.42, p = .02, d = .59 \). It is acknowledged that “number of descending triples” is not a measure of heterogeneity; instead this analysis indicates that successful solvers were more likely to produce iterations of informative descending triples.
Chapter 8

Experiment 6: Manipulating Contrast Class Information across Axes

Table 8-4

Summary of Triple Generation Profiles across Success on the Task

<table>
<thead>
<tr>
<th>Measure of Heterogeneity</th>
<th>Non solver</th>
<th></th>
<th>Solver</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
<td>SE</td>
</tr>
<tr>
<td>Constant positives</td>
<td>3.64</td>
<td>0.25</td>
<td>4.11</td>
<td>0.31</td>
</tr>
<tr>
<td>Posvars</td>
<td>0.97</td>
<td>0.21</td>
<td>1.00</td>
<td>0.23</td>
</tr>
<tr>
<td>Number of triples receiving negative feedback</td>
<td>5.44</td>
<td>0.24</td>
<td>4.79</td>
<td>0.24</td>
</tr>
<tr>
<td>Negtypes</td>
<td>2.69</td>
<td>0.23</td>
<td>2.07</td>
<td>0.28</td>
</tr>
<tr>
<td>Triptypes</td>
<td>4.11</td>
<td>0.23</td>
<td>3.64</td>
<td>0.34</td>
</tr>
<tr>
<td>Descending triples</td>
<td>1.94</td>
<td>0.30</td>
<td>2.93</td>
<td>0.26</td>
</tr>
</tbody>
</table>

8.4 General discussion

The purpose of Experiment 6 was to test empirically the contrast class account of DG facilitation by manipulating the information available to participants. The experiment offered support for the contrast class account of the DG facilitation effect and close analysis of participants' triple generation profiles failed to provide much support for competing accounts. The goal complementarity approach suggests that it is the production of posvar triples which is most closely associated with success on the task; the data in Experiment 6 do not support this account. Neither do the data support the information quantity approach, as there are differences in performance across conditions despite the fact that the number of triples being elicited was kept constant. The triple heterogeneity account suggests that solvers will differ from non-solvers in the range of triples produced, a prediction which is again not supported by the present dataset.

However, it was previously suggested that triple heterogeneity may be instrumental in promoting success on the DG version of the task in that a wide search of the hypothesis space leads to the generation of a contrast class triple which, in turn, leads to task success. According to this view triple heterogeneity is seen as a useful preliminary stage in successful hypothesis testing. Some support for this idea is
Experiment 6: Manipulating Contrast Class Information across Axes

provided by the data analyses for Experiment 5 and 6. In both of these experiments participants were provided with contrast class information (albeit of varying utility) with the effect that they did not engage in an initial search of the hypothesis space, but were instead lured into the testing of a MED hypothesis informed by the initial exemplar. It could be argued, therefore, that the findings from both Experiment 5 and Experiment 6 offer a level of support for the triple heterogeneity theory of DG facilitation if a wide search of the hypothesis space is seen as an important first step in producing a potentially informative contrast class triple.

The findings of Experiment 6 are readily explained by the Extended Iterative Counterfactual Model (E-ICM) presented in Chapter 7. Experiment 6 demonstrated that it is not DG instructions per se which promote success on the task, but rather that it is useful contrast class information that differentiates solvers from non-solvers on the task. It was also suggested that standard DG instructions promote the production of these types of triples. The data from both Experiment 5 and 6 shows that DG facilitation can be hindered if the discovery of this useful contrast class information is impeded by presenting participants with a logically correct but psychologically non-useful MED triple. The E-ICM shows that participants initially generate and test a DAX hypothesis using iterative counterfactual testing until they have reached a criterion level of confidence in it, and then switch to the discovery and testing of the MED hypothesis. The model suggests that an initial MED hypothesis is based on the participant’s current DAX hypothesis: that knowing that the MED rule is different to the DAX rule they generate a MED hypothesis which is seen as “opposite” to the DAX rule. Using a similar strategy to that of the iterative counterfactual testing of the DAX rule they select a single property of the DAX rule and use this to generate a contrasting property – and therefore a triple which is an instance of this MED hypothesis. Thus, it would seem that one prediction of the extended ICM is that when asked for the opposite of a complex number rule (e.g., “even numbers ascending by two”) participants should produce a rule which varies on a single axes only (i.e., “even numbers descending by two”, “odd numbers ascending by two”, or “even numbers ascending by three”). The final study to be presented in this thesis tests this idea.
Chapter 9

Experiment 7: Exploring Contrast Class Generation for Presented Number Rules

9.1 Introduction

This work presented in this thesis so far has explored contemporary accounts of DG facilitation and has concluded that no single previously published account can adequately explain the effect, although it has been suggested that the triple heterogeneity account may be subsumed within the contrast class theory that has been developed as part of the present research programme. Chapter 7 described a study which aimed to explore the contrast class account of the DG facilitation effect and went on to present an extension of Oaksford and Chater's (1994) iterative counterfactual model (ICM) so as to demonstrate how contrast class ideas could be incorporated within the ICM so as to explain of the performance benefits that arise from DG instructions.

The extension to the ICM proposed that DG instructions facilitate rule discovery by encouraging the production of a MED triple based on the hypothesised DAX triple. It was suggested that due to the lack of direct information about the MED rule, two pieces of information are exploited by participants to allow them to form the basis of an initial working MED hypothesis: (1) their knowledge that the MED rule is different to the DAX rule; and (2) their current best hypothesis as to what the DAX rule might be. It was suggested that, in a similar manner to the counterfactual strategy utilised to generate and test the DAX hypothesis, a single property of this DAX hypothesis is selected and a triple is generated that varies from the DAX hypothesis on this axis only. This triple is then tested and if feedback is DAX the reasoner is made aware that their posited DAX hypothesis is erroneous and returns to testing the DAX hypothesis. If, however, feedback is MED the reasoner adds this triple to the list of MED triples and moves into the MED sub-loop, in which a number of positive instances of the MED hypothesis are tested (Klayman & Ha, 1987, 1989). Once the reasoner has generated sufficient evidence to be confident in their MED hypothesis they move out of the positive sub-loop and perform a comparison of their DAX and MED hypotheses in which any perceivable common properties of the two hypotheses are searched for. If none are perceived the reasoner announces the two rules. If, however, a common property is perceived this violates the assumption that the two rules are different and
Experiment 7: Exploring Contrast Class Generation

the process of DAX, and then MED hypothesis generation and testing is restarted using this common property as a basis for testing of the DAX hypothesis. Thus, throughout the model the reasoner examines their posited rules for common properties to inform their hypothesis formation and testing by varying a single property at a time. Such a model presumes a consistency across DAX and MED hypothesis generation and testing which would seem to be psychologically plausible, while the emphasis on examining a single property at a time allows for possible cognitive constraints on the process of hypothesis generation and testing.

One prediction of the extended iterative counterfactual model as presented here is that the initial MED hypotheses would differ from DAX hypotheses on a single axis only, regardless of the number of properties present in a reasoner’s working DAX hypothesis. This notion is based on Oaksford and Stenning’s (1992) notion of contrast class information. They note that psychological contrast class information is different from the logical opposite of a set of information, arguing instead that the most likely contrast class member should be as similar as possible to the negated constituent. Oaksford and Stenning use the example “Johnny didn’t travel to Manchester by train” and point out that this sentence can be negated in one of three ways:

1) It was not Johnny who travelled to Manchester by train
2) That it was not to Manchester that Johnny travelled to by train
3) Johnny travelled to Manchester by some method other than by train.

They suggest that any of these three interpretations would be a valid contrast class to the sentence “Johnny travelled to Manchester by train”. Oaksford and Stenning argued that in this real world example it is assumed that people will use a combination of world knowledge (a person would be unlikely to travel by ship from London to Manchester) and intonation which directs the listener to the relevant constituent of the sentence to disambiguate where the negated constituent should lie. Note that the negations used to illustrate the contrast class to “Johnny travelled to Manchester by train” only negate a single constituent of the original sentence; it is this aspect of contrast class information which is employed in the extended iterative counterfactual model when suggesting that the reasoner seeking the opposite of the DAX hypothesis will vary it along one axis only.

One test of the extended iterative counterfactual model would, therefore, be to examine whether participants who are asked for the “opposite” of a number rule containing two or more constituents would, in fact, vary only a single constituent of that rule. During the course of this thesis a large body of data has been collected in the SG condition concerning incorrect initial rule announcements made by participants. It

- 119 -
was decided, therefore, that to improve the validity of this study it would be valuable to use these rule-announcement data as a basis for the number rules to be tested. It was predicted that when asked for the opposite of a number rule with two or more properties, participants would vary only one of these properties.

9.2 Method

9.2.1 Participants

One hundred and thirty-six staff and students from the University of Derby took part in the study on a voluntary basis. None had received any teaching relating to reasoning or logic.

9.2.2 Design

An independent measures design was used in which each participant was asked to state the opposite to one of the five number rules derived from a review of the incorrect rule announcements made in the single goal (SG) conditions of the previous experiments reported in this thesis that employed the standard "ascending numbers" target rule (i.e., Experiments 1 to 3). The rule allocated to each participant was randomly selected from the set of five available, with the constraint that thirteen participants were given the "numbers rising by two" rule for every one participant presented with each of the other rules (see Materials section for further explanation).

9.2.3 Materials

The rules to be used were generated by reviewing the rules produced by participants in the SG condition in previous studies in this thesis. A total of 90 participants had received SG instructions with a target rule of ascending sequences, 48 of whom had announced an over-restricted rule (a further seven had produced other incorrect rules). These over-restricted rules were of the form:

i. Rising by two

ii. (Going up) in two times tables

iii. Going up in equal intervals

iv. Even numbers going up.

v. Even numbers going up in twos.

These rules were announced by participants in an approximate ratio of 13:1:1:1:1. If the extended ICM is an accurate reflection of people’s DG hypothesis testing strategy
then these rules represent participants’ working DAX hypotheses as they move on to
the generation and testing of the MED rule. Although the primary aim of the present
study was to examine whether participants varied complex rules on either single or
multiple axes, it was decided that given the context of the experiment the most valid
approach would be to mirror the incorrect DAX hypotheses produced in the SG
condition. For this reason these rules were presented in the above proportions to
participants in the present experiment.

9.2.4 Procedure

Participants were tested individually and verbally. Each participant was given a
number rule from the above pool, and was asked what they thought the “opposite of
the rule” would be. Answers were scored on a simple tick sheet which recorded which
the rules they were asked about and possible contrasting rules. For rules (i) to (iv)
there were three possible alternatives which were expected to be generated; two which
varied from the given rule on a single axis, whilst the third varied on two axes. For
example possible alternatives to the first rule (rising by two) are:

a. Descending in twos (single axis)

b. Rising by a number other than two (single axis)

c. Descending by a number other than two (two axes)

There were seven possible “opposites” to rule (v); three varied on a single axis, three
on two axes and one rule mentioned the opposite of all three components in the
example rule:

a. Odd numbers going up in twos (single axis)

b. Even numbers going up in other than twos (single axis)

c. Even numbers going down in twos (single axis)

d. Odd numbers going down in other than twos (two axes)

e. Even numbers going down in other than twos (two axes)

f. Odd numbers going down in twos (two axes)

g. Odd numbers going down in other than twos (three axes)

9.3 Results

It had been predicted that when asked for the opposite of a number rule with two
or more properties, participants would vary only one of these properties. Of the 136
participants in the study only a single idiosyncratic answer was given (zero) and this was removed from the dataset. No participant produced a rule which varied on more than two axes, although such an eventuality was possible in rule (v). To analyse the data a comparison was therefore made of the number of participants who generated a rule which varied on a single axis against the expected frequency of 50% predicted by chance. The analysis showed that 89% of the participants generated a "single axis" variation, which a binomial test indicated was highly reliable, \( p < .001 \).

### 9.4 Discussion

The results reported here unequivocally support the notion that when asked for the opposite to complex number rule, participants focus on a single axis and negate that single axis only. This finding is in line with a prediction of the extended ICM presented in Chapter 7. According to extended ICM when asked to find two rules with the DG paradigm a participant initially focuses on discovering the DAX rule; although this in not made explicit by the model this is assumed to be the case because the reasoner has an exemplar triple on which to base their initial hypothesis, whilst they have no information about the MED rule. Once confident of their DAX hypothesis the reasoner then turns their attention to the MED rule. Unlike the DAX rule they do not have a MED exemplar (in the standard form of the task), and they must therefore induce the MED rule from information that is available to them: they know that MED is different from DAX and they think that DAX is (for example) "numbers rising by two". It was argued that they would therefore induce a MED rule which exploited these two pieces of information and generate a MED hypothesis which differs from the DAX hypothesis as much as possible (i.e., produce a rule which is "opposite" of DAX). Based on Oaksford and Chater’s (1994) ICM - which is concerned with exploring one common feature at a time - it was suggested that rather than generate the logical opposite of their "number rising by two" DAX hypothesis participants would instead vary only one of the two possible axes.

It was suggested in Chapter 5 that when seeking to find the "opposite" of their DAX rule it is likely that participants will vary the axis which has a logical opposite (i.e., ascending versus descending), rather than the more fuzzy "by two". A second analysis was therefore performed which focused on those participants whose presented rules contained both a "fuzzy" and a "non-fuzzy" component. This analysis showed that 83% of "opposite" rules announced varied the non-fuzzy component of the rule, which was significant with a binomial test, \( p < .001 \). By contrast, when participants were asked for the opposites of rules containing two "non-fuzzy" components (e.g., "even numbers going up") it emerged that the axis which was
varied was split more evenly (46% versus 54%). However, there were very small numbers in this latter group, making it difficult to draw conclusive inferences.

In summary, the data presented here offer support for the extended ICM model (Chapter 7) in two ways. First, when asked to produce the opposite of a complex number rule it is evident that participants will vary only a single axis of the rule. Second, when presented with rules which vary on both “fuzzy” and “non-fuzzy” axes it is clear that participants will produce many more rules in which the “non-fuzzy” aspect is varied rather that the more difficult “fuzzy” axis.
Chapter 10

General Discussion

10.1 Introduction

The purpose of this thesis was to explore contemporary accounts of DG facilitation in Wason's (1960) rule discovery task in an attempt to discriminate between them with a view to providing an overarching theory of the DG effect. It was also hoped that a clearer understanding of why DG instructions improved performance would go some way toward explaining the poor performance in the SG variant of the task. In this final chapter an overview of the experiments performed in this thesis is provided, along with a description of the development of a novel "contrast class" account of the DG facilitation effect. In addition, it is shown how Oaksford and Chater's (1994) iterative counterfactual model (ICM) can be extended to provide a process model which accounts for the data presented in this thesis and the DG facilitatory effect more generally. An evaluation of this model is also presented. The Extended-ICM (E-ICM) is then situated more broadly within ideas relating to the philosophy of science, and it is argued that the model promotes the use of a range of hypothesis testing strategies. More specifically it promotes the use of a simplified form of Bayesian hypotheses testing in which it is not necessary for the reasoner to engage in *a priori* estimates of the diagnosticity of proposed tests. Finally, suggestions are made for further studies designed to explore the ideas implicit in the E-ICM and hypothesis testing strategies more generally.

10.2 Overview of Experiments

10.2.1 Testing Evans' (1989) Positivity Bias Theory of DG Facilitation

In Experiment 1 Evans' (1989) positivity bias theory was explored. According to this theory DG facilitation is attributable to the linguistic labelling of the feedback given to participants rather than the specific requirement to discover two rules. Evans claims that relabelling the negatively valenced "does not fit" feedback as "MED" overcomes a preconscious tendency to attend to positive information and discard negative information. In Experiment 1 these assumptions of the positivity bias theory were tested by independently manipulating instruction mode (SG versus DG) and linguistic labelling of feedback ("fits/does not fit" versus "DAX/MED"). Contrary to
the predictions of the positivity bias theory no effect of linguistic labelling of feedback was apparent, with participants being more successful in the DG condition regardless of the form of feedback. In terms of solution success, then, Experiment 1 appeared to undermine the positivity bias account of DG facilitation.

Analysis of the triple generation profiles in Experiment 1 allowed observations to be made regarding other accounts of DG facilitation. Some support was gained for triple heterogeneity theory (Vallee-Tourangeau et al., 1995), with DG instructions promoting a wider variety of negtype triples, more triples receiving negative feedback and more posvar triples. The association of posvar generation with success on the task also supported Wharton et al.’s (1993) goal complementarity theory of DG facilitation. Furthermore, the greater numbers of triples that were produced by those in the DG condition lent some support to Wharton et al.’s (1993) information quantity theory of DG facilitation.

Overall, then, Experiment 1 gave rise to a degree of support for all contemporary theories of DG facilitation with the exception of the positivity bias account. Careful exploration of the data, however, revealed a previously unremarked phenomenon: that it was the production of a descending triple rather than the intuitively more informative posvar triple which was most closely associated with success on the task. This novel finding was intriguing in that contemporary theories of DG facilitation have generally been silent on the role of negative triples in the task. Moreover, if the effect was found to be robust then a theoretical account for the finding would seem to be essential.

10.2.2 Testing Wharton, Cheng, and Wickens’ (1993) Information Quantity Account of DG Facilitation

Experiment 2 was designed to explore (Wharton et al., 1993) information quantity account of DG facilitation. This theory is based on two premises: that solvers produce more triples than non-solvers before announcing the rule (cf. Wason, 1960), and that participants given DG instructions reliably produce more triples than those given SG instructions (Gorman et al., 1987). Testing the information quantity account therefore involved manipulating the number of triples participants were required to produce across instruction mode. Thus participants were given either DG or SG instructions and were required to produce 6, 12 or 18 triples. Results of this study were equivocal with regard to the information quantity theory. First, contrary to information quantity theory predictions, performance in the DG condition was superior to that in the SG condition. Second, however, those participants asked to produce 12 triples were more successful than those in the 6 and 18 triple conditions.
This latter observation tends to indicate that although the relationship between number of triples produced and success is not linear, there is an optimum number of triples for success on the task. The observation that participants who were asked to produce 12 triples had the highest rate of rule discovery concurs with previous reports in the literature suggesting that the mean number of triples produced by solvers deploying "natural" hypothesis testing strategies is between 10 and 12 (e.g., Farris & Revlin, 1989a, Experiment 1; Gorman, 1987, Experiment 1; Vallée-Tourangeau & Payton, 2008). It would seem, therefore, that the number of triples generated by people may well have a role to play in successful hypothesis testing, but this role is more complex than suggested by information quantity theory.

In terms of other theories of interest, the data from Experiment 2 offered some support for both the triple heterogeneity theory (Vallée-Tourangeau et al., 1995) and the goal complementarity theory (Wharton et al., 1993). It was further noted that the presence of a contrast class triple was again associated with success on the task, although for all three accounts the evidence was not entirely convincing. Given the equivocal evidence for all theories of interest it was decided that none could be discarded and that all should be considered in further experiments, but with controls being put in place for the information quantity account by limiting triple production to 10 triples in order to mirror naturally successful hypothesis testing behaviour.

10.2.3 Testing Wharton, Cheng, and Wickens' (1993) Goal Complementarity Theory

Experiment 3 was concerned with direct testing of goal complementarity theory (Wharton et al., 1993). On initial inspection this theory seemed to be the most convincing in that it provides a clear rationale for both how and why DG instructions promote success on the task. The study systematically manipulated the relationship between the DAX and MED rules so that in some conditions they were no longer strictly complementary. Contrary to the basic tenet of the theory (i.e., that it is the mutually exclusive and exhaustive relationship between the two rules which promotes success) in one "non-complementary" condition performance was similar (and indeed marginally superior) to that in the standard, complementary condition. This finding was interpreted in line with contrast class theory, in that it was argued that in the non-complementary but facilitatory DG condition the sought-after MED rule provided clear contrast class information to the reasoner, whereas in the standard DG condition contrast class information provided by the MED rule was less clear cut.

A focussed follow-up study was conducted that further explored the goal complementarity theory using different rules. This study offered additional evidence
for contrast class ideas in that facilitation was greatest when MED feedback provided
the strongest contrast class cue to the DAX rule, while in the DG-4 condition where
the MED rule did not provide useful contrast class information performance was very
poor. It was again noted that while support for the triple heterogeneity theory (Vallée-
Tourangeau et al., 1995) was somewhat limited, data analysis did show some level of
support for this account, and therefore triple heterogeneity theory needed to be
considered in more detail in further experiments. It was suggested, however, that
whilst it was possible that a wide search of the triple-space was a useful preliminary
step in identifying contrast class triples (which in turn leads to success on the task)
such broad triple search may not in itself be the root cause of task facilitation.

It was at this stage, with support for the notion of contrast class information
playing a critical role in successful rule discovery, that it became important to identify
a theoretical explanation for the effect. It was noted that it might be possible for
Oaksford and Chater’s (1994) iterative counterfactual model (ICM) to go some way to
explaining the effect, although it was also acknowledged that the original model was
concerned with hypothesis testing using SG instructions and would therefore need to
be extended to include DG instructions.

10.2.4 Generalising the Contrast Class Theory to Other Rules

Evidence had now been accumulated suggesting that when the original
“ascending numbers” DAX rule was used in the DG paradigm then the generation of a
“descending” contrast class triple seemed to be the critical determinant of successful
rule discovery. It was therefore decided that before developing the contrast class
theory further it would be prudent to ensure that the effect was not merely an artefact
of the DAX rule being used. Experiment 4 was therefore a simple implementation of
the 2-4-6 task using SG versus DG rules and Gorman et al.’s (1987) more difficult
“three different numbers” rule. Analysis of the data showed DG facilitation with this
more difficult rule, and also demonstrated that using this alternative rule production of
a contrast class triple remained predictive of success on the task.

It was additionally shown that generating a wide range of triples was also
associated with success on the task, although modelling of the data showed that this
effect disappeared once production of a contrast class triple was controlled for. The
evidence for competing theories of DG facilitation seemed again to favour the contrast
class theory of the effect, although it was clear that triple heterogeneity theory should
not be discarded.
10.2.5 Testing the Contrast Class Theory of DG Facilitation

Evidence for both contrast class theory and triple heterogeneity theory presented up to this point had been a result of post hoc analyses of triple generation profiles, rather than a consequence of direct manipulations of available contrast class cues. To address this weakness it was decided to explore contrast class theory empirically. Contrast class theory proposes that it is the production of a triple which varies on the salient ascending/descending axis which promotes success on the task. The aim of Experiment 5 was therefore to test contrast class theory by providing MED information of varying utility to participants and examining the effect of the information on success rates.

Results provided strong evidence for contrast class theory in that those given useful (i.e., “descending”) contrast class information were significantly more successful than those given non-useful (i.e., “equal numbers”) contrast class information. Indeed, it was found that non-useful contrast class information impeded performance on this DG version of the task, with the result that success rates were similar to those normally reported using SG instructions (e.g., Tukey, 1986, Wason, 1960).

This study provided no evidence that varied triple testing was predictive of task success. However, this null effect was interpreted as indicating that participants in this study did not actually need to engage in hypothesis space searching as they had been provided with a contrast class cue in the first place. As such, participants were already at a more advanced stage of the hypothesis testing process than would typically be the case, and being at this advanced stage obviated the need for a broad triple search. Overall, then, it remains possible that under standard DG instructions a wide search of the hypothesis space is a key precursor to the generation of a useful contrast class triple.

10.2.6 Testing Contrast Class Theory by Manipulating Information across More than One Axis

In addition to providing evidence for the contrast class theory, Experiment 5 also led to a suggested extension to Oaksford and Chater’s (1994) ICM which illustrated how implementation of DG instructions would affect the model, and specifically why contrast class triples are generated. This model (see Figure 10.1) proposed that when generating a MED hypothesis, participants focus on a single attribute of their working DAX hypothesis so as to generate a MED hypothesis which varies on this axis only, leading to the generation of a positive example of this MED hypothesis. A prediction deriving from the model is that the most useful MED triple would be one which
varied from the DAX hypothesis on a single axis only. Experiment 6 empirically tested this idea by comparing the performance of participants given MED exemplars which varied on either single or multiple axes. As predicted, performance was best for those participants given exemplars with useful “single axis” contrast class information. *Post hoc* analysis of triple generation profiles again showed no evidence that varied triple testing was predictive of success on the task, which offered further support for the idea that a wide exploration of the triple-space may simply be a precursor to generating contrast class information rather than being an important predictor of success in itself.

10.2.7 Exploring the Generation of Contrast Class Information

The final experiment explored the idea implicit in the extended ICM (E-ICM) that after producing a plausible DAX hypothesis participants move to generating and testing a MED hypothesis which is formed by seeking the “opposite” of their DAX rule, and that this “opposite” will vary from their DAX rule on single axis. It was found that when presented with complex number rules the majority of participants generated opposite rules which did indeed vary on just one axis, and that this axis was one for which a logical contrast is most easily brought to mind.

10.2.8 Summary of Empirical Evidence

The empirical evidence presented in this thesis appeared to suggest that none of the contemporary theories of DG facilitation presented in the introduction provided a wholly convincing account of the phenomenon. As a result, and in the light of new effects noted in the reported data analyses, an extension to Oaksford and Chater’s (1994) ICM was proposed. This new model, termed the extended iterative counterfactual model (E-ICM), appeared to provide a comprehensive explanation of the DG effect, and aspects of the reported research attempted to test core predictions deriving from the model.

It was further proposed that Vallée-Tourangeau et al.’s (1995) triple heterogeneity account may have a role to play in explaining the DG facilitation effect in that a wide search of the triple-space may be a preliminary stage in generating a contrast class triple, which the E-ICM suggests is the critical determinant of success on the task. Although not previously discussed it is also possible that information quantity (Wharton et al., 1993) may play a similar role in the 2-4-6 task, in that the increased triple generation evident in both successful SG participants and DG participants may be an artefact of wide triple-space search and therefore underpins the generation of a contrast class triple.
The evidence presented in this thesis does not offer direct support to the notion that these alternative theories of DG facilitation can be subsumed within the E-ICM model which has been proposed, but neither does it exclude this possibility. However, while evidence for both the triple heterogeneity theory and information quantity theory appears to be less convincing than for the contrast class theory, some support for both is evident, both in this thesis and in the literature more generally (e.g., Vallée-Tourangeau et al. 1995; Wharton et al. 1993). As such these theories cannot be discarded and need to be included in any account of DG facilitation. The suggestion that strategies may change within a hypothesis testing session fits well with the work of Roberts (2000; 1993) and Siegler and Chen (2002), who view reasoning as involving a range of potentially overlapping strategies. This approach is taken in the present proposals relating to the E-ICM.

10.3 Contrast Class Theory and the Extended Iterative Counterfactual Model

The evidence presented in this thesis suggests that it is the production of a contrast class triple which is of critical importance in determining success on Wason's rule discovery task. For contrast class theory to be a convincing account of DG facilitation it is important to be able to explain both why DG instructions promote the generation of a contrast class triple and why this is not the case with SG instructions. A clear rationale for why production of a contrast class triple promotes success on the task is also required.

It is suggested that DG instructions promote generation of a contrast class triple by the following mechanism. Following the generation and testing to satisfaction of the DAX hypothesis, the participant turns their attention to discovering the MED rule. Unlike the DAX rule (for which an example triple is given) the participant has no explicit information on which to base their MED rule. They therefore use the implicit information that the DAX and MED rules are different and that the DAX rule is something like "numbers rising by 2". Oaksford and Stenning (1992) have shown that when considering the negation of a complex statement participants generally negate only a single aspect of the original statement. They suggested that in everyday life people use a combination of world knowledge and intonation to disambiguate which constituent of the sentence should be negated. In the 2-4-6 task such cues are not available to the participant, instead they must use information contained within the DAX rule to cue them to a relevant negation. The most common initial DAX rule is of the form "numbers rising by two" which contains two axes which may be negated: "ascending" and "by two". It is much easier to access the logical opposite of
“ascending” than "by two”, and it is therefore very likely that this would be the component of the statement which would be negated; indeed evidence from Experiment 7 supports this notion. It is therefore suggested that it is by this mechanism that DG instructions promote the generation of a contrast class triple. This mechanism relies wholly on the direct instruction to seek a second rule, and it is therefore clear that SG instructions would not directly promote generation of a contrast class triple.

The second point which needs to be clarified for the contrast class theory to be a convincing account of DG facilitation is to show how the generation of a contrast class triple promotes success on the task. In the early stages of this thesis it was argued that such a triple increases the salience of the “ascending” aspect of the DAX rule. A more focussed explanation was developed however, in which it was shown how contrast class theory could be incorporated into the most psychologically plausible account of hypothesis testing behaviour in the 2-4-6 task: Oaksford and Chater’s (1994) iterative counterfactual model (ICM). The revision to the ICM suggests that the processes by which the DAX hypothesis is generated and tested remain in place, that is, participants extract a property from the initial DAX exemplar (e.g., “even numbers”), generate a complement to this (e.g., “odd numbers”) and then produce a positive instance of this complementary hypothesis. DAX feedback informs the participant that their initial DAX hypothesis is incorrect, so the instance is added to the list of DAX instances and another common property extracted which is tested in the same way. If, however, MED feedback is given, the participant knows their DAX hypothesis is plausible and they move into the positive sub-loop in which examples of the DAX hypothesis are generated and tested, until a preset criterion level of confidence is reached.

It is at this point that the extension to the model is proposed; rather than announcing their DAX rule the DG participant now moves to the generation and testing of the MED hypothesis. As noted above, it is suggested that the participant uses information from DAX to generate a MED hypothesis which varies from the DAX hypothesis on a single axis. This newly generated MED hypothesis is then tested using a positive test strategy (see Figure 10.1). If DAX feedback is given the participant is alerted to the over-restricted nature of their DAX hypothesis and therefore returns to the generation and testing of DAX hypotheses. If, however, MED feedback is given the participant moves into the MED positive sub-loop and tests instances of their MED rule until a criterion level of confidence in their MED hypothesis is reached. The participant now compares their DAX and MED hypotheses, seeking common properties. If a common property is found the participant
will seek to relieve the tension caused by the sharing of a commonality between the
two hypotheses which are perceived to be different by returning to the testing of the
DAX hypothesis, generating a complement to it and passing through the stages of
DAX and MED testing until no further common properties are perceived. At this point
the DAX and MED hypotheses are announced and the study is over.

It is clear that this approach will promote a more diligent search of the problem
space in terms of both the number and type of triples produced as the participant
moves through the stages of the model and therefore both the information quantity
(Wharton et al., 1993) and triple heterogeneity (Vallée-Tourangeau et al., 1995)
theories can be subsumed within this model. This is an important point as although
strong evidence was not found for either account in this thesis, both theories have
previously found support in the literature (e.g., Tukey, 1986; Vallée-Tourangeau et al.,
1995; Wharton et al., 1993). Both theories have been characterised as being
descriptive rather than explanatory (Gale & Ball, 2005; 2006), and it is encouraging to
be able to provide a theoretical mechanism by which such effects may be explained.

10.4 Evaluation of the Extended Iterative Counterfactual Model

This thesis has proposed an extension of Oaksford and Chater’s (1994) iterative
counterfactual model (ICM) to incorporate DG instructions. It has been proposed that
when presented with DG instructions reasoners typically initially engage in
discovering the DAX rule and only when they are reasonably confident of its veracity
do they move to exploring the MED rule. The extension to the model suggests that the
strategies used to generate a MED hypothesis are similar to those used in generating
the DAX hypothesis. According to the ICM, when searching for a SG rule participants
extract a list of common properties from the list of DAX instances and select one of
these properties to produce a complementary hypothesis (H') which they then test
using a positive example. Feedback from this instance is used to inform either further
testing of the original DAX hypothesis (if feedback was “no”) or to development of a
refined DAX hypothesis (if the feedback was “yes”). Thus, generation and refinement
of the DAX hypothesis proceeds by focussing on single common properties of the list
of DAX exemplars.

The extension to the model proposed in this thesis builds on this strategy of
considering a single axis at a time by suggesting that MED hypotheses are derived
from DAX hypotheses by identifying a single property of the DAX hypothesis and
varying it to generate the MED hypothesis, which is then subjected to testing. The
suggestion that MED rule generation mirrors DAX rule generation is psychologically
plausible; it is likely that such a strategy would reduce the cognitive load which might be caused by using different strategies for DAX and MED hypothesis testing and generation. That there is a level of cognitive constraint in successful hypothesis testing behaviour can be adduced from the recent work of Vallée-Tourangeau and colleagues (Vallée-Tourangeau & Payton, 2008; Vallée-Tourangeau & Penney, 2005) who showed that when some of the mental load involved in the rule discovery task was externalised success rates in the SG version of the task rose. In addition to the empirical evidence which has been presented in this thesis supporting the notion that MED hypothesis generation is concerned with varying a single axis at a time, Tschirgi (1980) showed that when presented with hypothesis testing tasks in which multiple causative agents are imputed then reasoners adopt a strategy of “varying one thing at a time”.

For the E-ICM to be a convincing account of DG hypothesis testing strategies it is important that it is able to account for all of the effects reported in the DG literature. In Experiment 3 and its associated follow-up study although participants were given DG instructions the relationship between the DAX and MED rules was varied such that they were no longer strictly complementary. Contrary to the predictions of goal complementarity theory (Wharton et al., 1993) facilitated performance was still found in one of the non-complementary conditions. How, then, can the E-ICM account for this finding? It is proposed that initial DAX and MED hypothesis generation and testing remain as for the standard DG instructions. However, with the revision of the relationship of the rules it is now possible that participants will generate an instance of their MED hypothesis (Box 17, Figure 10.1) which they are told “fits” neither the experimenter’s DAX nor MED rule. They must now reject their MED hypothesis (Box 22, Figure 10.1), but contrary to the process when standard DG instructions are used, they can no longer add this instance to the list of DAX instances. In fact, this iteration of triple generation and feedback provides very little information about their DAX hypothesis. It is therefore likely that the reasoner will now return to their DAX hypothesis and use this to generate a new MED hypothesis (see Figure 10.1). By this process the definition of the MED rule (and by default the DAX rule) can be refined. The enhanced performance using DG-3 instructions can be attributed to the fact that in the standard DG condition triples which are not necessarily descending in nature (e.g., “three equal numbers”) will still be given MED feedback, and that this feedback may cause the reasoner to move prematurely into the MED positive sub-loop. Positive testing of a non-descending hypothesis will not reveal the over-restricted nature of the MED hypothesis (which, of course, should be “any non-ascending triple”) which in turn may lead to erroneous DAX rule announcement as demonstrated in Experiment 5. The enhanced performance in the DG-3 condition can therefore be attributed to
Figure 10-1

*Extended Iterative Counterfactual Model for Non-Complementary Rules*
the “descending” definition of the MED rule preventing premature movement into the positive MED testing sub-loop.

The very poor performance on the DG-4 version of the task can also be explained by contrast class theory (and by extension the E-ICM). When using DG-4 instructions the relationships between the rules was varied such that in addition to allowing triples to be DAX, MED, neither or both, information on the MED rule was no longer informative regarding the DAX rule. For example in Experiment 3 the MED rule was “odd numbers”, which does not contrast with the “ascending numbers” DAX rule. Indeed, by introducing a definition of MED which varies from DAX in the type of rule used (i.e., object regularities versus relational regularities; see (Cherubini, Castelvecchio, & Cherubini, 2005) performance levels were reduced below those typically reported when using SG instructions. One view of performance rates when using such DG-4 instructions is that the participant is effectively engaged in tackling two SG tasks simultaneously. A second (and perhaps more contentious) view is that participants may now be engaged in a disjunctive reasoning task, the type of reasoning task which Evans (1989) views as the most difficult of all. Indeed, in his review of the literature on the THOG problem (Wason & Brooks, 1979) Evans (2007) suggests that this difficulty may be due to the fact that participants are required to reason simultaneously about two mutually exclusive hypotheses which violates the singularity principle of hypothetical thinking. Evans’ explanation fits very neatly with the E-ICM account of the DG-4 success rates reported here.

10.5 Testing the Extended Iterative Counterfactual Model

The final three chapters of this thesis tested some of the predictions derived from the E-ICM model and provided further evidence for the contrast class theory of DG facilitation. Stronger testing of the model involving the collection of direct evidence about the hypothesis testing strategies of reasoners would be a useful next step. This, however, could be difficult. Collection of verbal protocols is one possibility, although questions have been raised in the literature about the validity of such verbal data. For example, Gagné and Smith (1962) showed that requiring participants to verbalise their thoughts when working on the Tower of Hanoi problem improved performance by reducing the number of moves taken to reach the solution when compared to a silent control group. On the other hand, Schooler and Engstler-Schooler (1990) showed that verbalisations hindered performance on insight problem solving, an inhibitory effect which they referred to as “verbal overshadowing”. Schooler and Engstler-Schooler suggested that such verbal overshadowing may arise because processes that are not easy to verbalise are involved in developing a solution to insight problems, and
therefore the requirement to verbalise can lead to participants focussing on easy to verbalise but not necessarily helpful processes. In a more recent study Fleck and Weisberg (2004) explored the use of verbal protocols in Duncker's (1945) candle problem and found no difference on a number of measures of success between those required and those not required to verbalise their thoughts. In a comprehensive review of verbalisations on the performance of experts on a variety of tasks Ericsson (2006) concluded that provided careful instructions were employed it is possible to elicit valid non-reactive reports of participants' thoughts during a cognitive task (see Ericsson & Simon, 1980 for detailed instructions for eliciting valid verbalisations of the ongoing cognitive processes). Evidence regarding the validity of the use of verbal protocols during reasoning tasks is, therefore, contradictory, and if such a tool was used careful controls and instructions would need to be implemented to ensure the validity of any data collected.

A second approach that could be adopted - and one that is more in line with standard methodologies used in Wason's 2-4-6 task (e.g., Vallée-Tourangeau & Payton, 2008; O. Vartanian, C. Martindale, & J. Kwiatkowski, 2003b; Wason, 1960) - would be to ask participants to record their hypotheses and reasons for test choice on their record sheet. Such an approach would allow closer exploration of why particular triples are chosen, and to assess more directly which hypothesis is being tested. When this methodology is employed, however, examinations of these reasons show that rather than providing reasons for a triple choice participants tend to respond with their current hypothesis. Gorman et al. (1987) adopted a slightly different approach and asked participants to predict whether triples would be DAX or MED before receiving feedback, while Tweney et al. (1980) asked participants to label each triple as either confirmatory or disconfirmatory. While useful data can be, and, indeed, has been, collected using such instructions, they do not capture the second-to-second processing which may be occurring during engagement with the task. One possibility which may be useful would be to ask participants to record which rule they are trying to find with each triple generated. E-ICM would predict that participants would initially test DAX hypotheses and then move to MED hypothesis testing. Post hoc examination of participants' protocols would allow testing of the idea that if a participant states they are looking for a MED rule and receive DAX feedback their next triple should be an attempt to redefine the DAX rule in the light of the extended list of DAX instances.

Another alternative, based in a Bayesian analysis of the task, would be to ask participants to state their current hypotheses and also to request them to place confidence ratings in these hypotheses. According to both the Bayesian philosophy of science and Oaksford and Chater's (1994) ICM confidence in DAX hypotheses should
gradually rise as participants generate examples of these hypotheses whilst in the positive sub-loop. It would be expected that once the participant is reasonably confident in their DAX hypothesis they would move to testing of the MED rule; confidence in this would initially be low, but again would rise as they worked through the MED positive sub-loop. Such a process would inevitably be rather complex and would need careful instruction and implementation to ensure that intrusive requirements do not affect performance on the task, as well as careful coding of confidence ratings relating to the testing of hypotheses. It is of note that Tukey (1986) used a similar methodology in his investigation of hypothesis testing strategies, but the limitations in his dataset did not allow conclusions to be drawn regarding participants’ confidence in their hypotheses. In addition, Tukey’s participants were given SG instructions, and while data regarding the number of triples to first announcement were not reported it is likely that in common with many other studies employing SG methodologies his participants may have “spoken too soon”.

10.6 DG Facilitation and the Philosophy of Science

The literature review earlier in the thesis (see Chapter 1) introduced some ideas from the philosophy of science regarding the method by which scientific discovery advances. In the original version of the 2-4-6 task Wason (1960) set out to discover if participants adhered to the then contemporary norm of scientific testing, that of falsification (Popper, 1959). Wason designed the task such that the initial exemplar would induce an overly-restricted rule (Wetherick, 1962), and that the only method of discovering that this rule was erroneous was by attempting to disconfirm it. Participants in Wason’s original study appeared to demonstrate a maladaptive hypothesis testing strategy; that is, the majority of participants engaged in what Wason characterised as an “enumerative” strategy rather than the more informative “eliminative” strategy.

Wason’s original interpretation of performance on the task as being due to confirmation bias pervaded the literature surrounding the task for some time, with attempts to improve performance being focussed on trying to eliminate confirmatory behaviour in participants with generally mixed results (e.g., Gorman & Gorman, 1984; Gorman, Stafford, & Gorman, 1987; Kareev et al., 1993; Mynatt, Doherty, & Tweney, 1977; Mynatt et al., 1978; Tweney et al., 1980). It was shown that while instructional manipulations were successful in promoting generation of disconfirming triples this did not always translate into improved success rates (e.g., Tweney, Doherty, Womer, Pliske, Mynatt, R.Gross et al., 1980). In contrast, participants in Gorman et al.’s (1992b) study did show an improvement when he did not provide feedback on
participants’ rule announcements: instead he instructed participants to decide for themselves if they were right. Gorman interpreted this as showing than once participants could no longer appeal to the experimenter to confirm or disconfirm their hypothesis (which he noted was much more powerful than producing triples to evaluate a hypothesis) then confirmation became a much stronger heuristic for successful solving.

In its original form the 2-4-6 task was, therefore, seen as modelling scientific discovery, for which the strongest methodology was to use a Popperian falsificatory strategy (Popper, 1959), and the evidence suggested that participants were unable to adopt such a strategy in the laboratory. The DG variant (Tweney, Doherty, Womer, Pliske, Mynatt, R.Gross et al., 1980), however, promoted much higher success rates on the task despite its close similarity to the original SG variant. Earlier in the thesis (see Chapter 1) it was argued that this was because the DG variant changed the demands of the task such that rather than a Popperian falsificiationist strategy being the most valuable the demands of the task could now be couched more naturally in a Bayesian form (Howson & Urbach, 2006) in which two competing theories are pitted against each other. Note, however, that a Bayesian approach can be taken to the SG form of the task. Recall that according to the Bayesian approach evidence is collected regarding a hypothesis, with belief in this hypothesis varying according to the strength of the evidence. When couched in Bayesian terms, however, each triple that is tested is assessed on the information gain from the test rather than being a simple falsification. The expected value of the information gain is, however, a rather complex calculation; for each test the range of possible outcomes needs to be known (in this case whether the triple “fits” or “does not fit” the rule) and also the probability of each outcome if the hypothesis is true or false. These data allow the calculation of the diagnosticity of each possible outcome using the following formula:

$$\frac{p(D/H)}{p(D/\sim H)}$$

In non-algebraic form this can be expressed as the probability of observing the outcome (datum) D given the hypothesis H is true, divided by the probability of observing the outcome D given that the hypothesis is not true. The result of this calculation informs the reasoner how much to change their belief in H should the test produce the outcome D. Clearly the test must be selected before the outcome is known so the expected outcome is influenced by the a priori probability of each outcome which can be derived from estimates of
Along with the *a priori* belief in H. By calculating these values it is possible to calculate which test is informative and discern when further testing is unlikely to yield additional information. A Popperian approach, however, only requires the reasoner to judge whether the result of a test will falsify their hypothesis, that is,

\[ p(H / -D) = 0 \]

when applied to the focal hypothesis.

Given the comparative simplicity of the calculation for a Popperian approach it would be expected that this would be the strategy of choice for the 2-4-6 task. However, while the calculation appears to be relatively simple the difficulty of this approach seems to be in producing a triple which would falsify a current hypothesis (Poletiek, 1996a). On the other hand, given the difference in the complexity of the calculations required for a Bayesian approach it is unlikely that reasoners would be able to deploy such an approach to the task (cf. Klayman, 1995). Tukey (1986) offered some support for the idea that a non-Bayesian approach is taken. In his study he asked participants to ascribe labels to triples they had generated in terms of their role in testing a current hypothesis. He found that very few triples were generated with a view to increasing certainty in a hypothesis, which Tukey interpreted as indicating a non-Bayesian approach to the task. Note, however, that the "assess" label used to denote a Bayesian approach referred only to "increasing" certainty that a hypothesis (rather than an alternative) is correct, a label which is perhaps more confirmatory than Bayesian in its connotations. Similarly, if the ICM (M. Oaksford & N. Chater, 1994) is taken as reflecting hypothesis testing behaviour then it could be argued that the inclusion of the positive sub-loop in the model suggests that reasoners do not engage in Bayesian testing; information collected at this stage of the testing process has a low diagnostic value in terms of the hypothesis being true. Note, however, that reasoners may be attempting a Bayesian approach as they are collecting information which is used to adjust the strength of their belief in their hypothesis; it may be that reasoners naturally adopt a Bayesian approach but the complexities of the calculation prevent the approach from being fully implemented. The positive sub-loop of the ICM also indicates that participants do not use wholly Popperian principles to test their focal hypotheses as there is no attempt in this phase of the hypothesis testing cycle to falsify
their current hypothesis, rather they collect evidence which supports it, even though the initial counterfactual stage of the process is falsificatory in nature.

The ICM therefore suggests that in the initial stages of hypothesis testing a Popperian approach is taken. A hypothesis is adduced from a list of instances which conform to the DAX rule and is tested, initially using a counterfactual strategy (which also leads to the generation of new hypotheses if the working hypothesis is disconfirmed). Once a plausible hypothesis is produced the participants then move to the positive sub-loop in which the non Popperian positive test strategy (Klayman & Ha, 1987, 1989) is used to generate instances of their hypothesis until a preset level of confidence is reached when the rule is announced. It is likely that in the SG form of the task it would be this testing of triples in the positive sub-loop which led to Wason (1960) identifying the "enumerative" testing strategy. The positive sub-loop is an interesting aspect of this account of people's SG hypothesis testing strategy in that it not only presents hypothesis testing in the SG task as possibly having Bayesian elements but it also mirrors the real life hypothesis testing strategies as reported by Mitroff (1974) and Tweney (1985). Both Mitroff and Tweney note that in the initial stages of research, scientists tend to try to confirm their hypotheses before moving to a falsificatory strategy. It might be that when the SG variant of the 2-4-6 task is employed participants disengage from the task before moving into the disconfirmatory stage, although it is unclear why this might be so.

In its DG variant the task can be characterised as facilitating a Bayesian approach in that two hypotheses are considered and evidence for these two hypotheses is collected and considered. According to the E-ICM success on the task is not due to negative testing of the DAX hypothesis, rather it is a result of positive testing of the MED hypothesis. According to the E-ICM evidence regarding the DAX hypothesis is collected before moving on to explore the MED hypothesis. The MED hypothesis is derived from the DAX hypothesis and, as such, can be seen as being related to it. While it is true the DAX and MED hypotheses are not competing in a true Bayesian sense, the fact that two hypotheses are related does allow for data to be collected and evaluated in the light of both of the hypotheses; that is, a priori calculations of the diagnosticity of possible tests do not need to be carried out. Information gained from the each test is of symmetric value to the two hypotheses under consideration, thus the demands of the DG variant may promote a simplified form of Bayesian hypothesis testing.

The interpretation of DG instructions as transforming a Popperian task to one that is Bayesian, together with the greatly enhanced performance in the DG variant, tend to suggest that a Bayesian approach to scientific discovery may be more useful.
than a strategy of falsification. However, this observation should be tempered by the knowledge that most studies employing Wason’s 2-4-6 task have used lay participants rather than scientists (see however Mahoney & DeMonbreun, 1977 for a study in which the performance of scientists and ministers is compared), and there remains the issue of ecological validity. A scientist engaged in research is aware that any published work will be open to peer review and testing by contemporaries, and it is therefore likely that they will be careful to test any hypothesis thoroughly before publishing. It would seem likely in such circumstances that thorough testing and evaluation would be likely to include elements of falsification. In this respect it is reasonable to suggest that although aspects of the 2-4-6 task mirror scientific testing, the lack of context in which scientific progress is made both in social terms and in the knowledge base which surrounds scientific discovery in the laboratory point to some shortcomings of it as a model. Thus, in this task at least, more note should perhaps be taken of Popper’s (1959) observation that scientific discovery is a special case of everyday reasoning; perhaps the demands of the 2-4-6 task more closely model everyday reasoning than scientific reasoning.

10.7 The “Hypothetical Thinking Framework” and Performance on the DG Task

In Chapter 1 Evans’ (2006) hypothetical thinking framework as related to the 2-4-6 task was discussed. According to this theory poor performance on the task can be attributed to a combination of heuristic and analytic processes. The participants’ initial erroneous hypotheses are caused by a heuristic relevance effect, in that information in the initial 2-4-6 exemplar is assumed to be relevant to the task (cf. Cherubini et al., 2005), with the effect that a sufficient but not necessary rule is generated. The analytic system does not intervene to revise the model generated by the heuristic system as the singularity principle leads to the consideration of a single hypothesis which is then evaluated according to the satisficing principle (Evans, 2007). In the case of the 2-4-6 task the initial hypothesis is found to have satisfied the demands of the task and is therefore accepted without consideration of further alternatives.

As was noted in Chapter 1 the dual process account is a weaker version of the “dual system” account of cognition which proposes that the two processes have differing evolutionary histories and originate in different neural substrates (Evans & Over, 1996; Stanovich, 2004). As such, it would therefore be expected that performance which appeared to be maladaptive by normative standards could be explained in terms of evolutionary advantage. This view is in line with the argument made earlier that the 2-4-6 task is more closely aligned with everyday reasoning than
scientific reasoning, as in evolutionary terms it is important to avoid costly errors (even at the expense of missing potentially useful benefits), whilst scientific discovery demands higher levels of reasoning which leads to solutions more closely approximating to the "truth" (Cosmides & Tooby, 1992). Evidence for this view comes from the work of Friedrich (1993), who in a wide ranging review of the hypothesis testing literature showed that when examined from the standard of the type of pragmatic logic required to negotiate through the social world it could be argued that the positive test strategy is superior to other strategies when identifying appropriate solutions in everyday life. According to Friedrich's (1993) Primary Error Detection and Minimisation (PEDMIN) analysis, people are more likely to be concerned with error reduction than truth detection, and that this concern shapes the questions people ask, and typically guides them towards sufficiency testing. According to the PEDMIN approach the fact that people tend to focus on false positives rather than false negatives is due to false positive errors being likely to be more costly than false negatives. This is illustrated by the example given by both Klayman and Ha (1989) and Friedrich (1993) that when buying a new car it is better to miss some perfectly good cars than buy one which is unreliable. The PEDMIN analysis then offers some support for the idea that hypothesis testing behaviour can be interpreted as resulting from evolutionary pressures to avoid errors, a view which can be subsumed within Evans' (2006) hypothetical thinking framework. Leyens, Dardonne Yzerbyt, Scaillet, and Snyder (1999) make a similar argument for the role of confirmation and disconfirmation and the development of in-group bias. Thus in terms of social cognition, arguments have been made for non-normative approaches having evolutionary advantages, providing indirect support for the dual process account of the task.

For Evans' (2006, 2007) hypothetical thinking framework to be a convincing account of hypothesis testing behaviours it also needs to account for performance using DG instructions. In Chapter 3 it was suggested that one way in which facilitated performance on the DG variant of the task could be accounted for would be that the requirement to find two rules would overcome the singularity principle. It was suggested that DG instructions lead to a more Bayesian approach as the two hypotheses are tested alternately, and participants' confidence in their two hypotheses varies as a function of the evidence collected. It was further suggested that this Bayesian approach would promote more triples of a greater variety as participants generated tests of their hypotheses.

It could be argued, however, that this latter view is somewhat undermined by the results of later experiments presented in this thesis which showed that the facilitatory
effect of DG instructions was apparent even when the number of triples required was held constant across instruction mode. Contrary to this argument is the notion that the requirement to generate a specific number of tests may have affected participants’ “natural” hypothesis testing strategies, leading them to construct tests more carefully if they felt testing was limited, and thereby promoting more diagnostic testing (in the DG conditions) or persistence in the positive test sub-loop (in the SG condition) after reaching their criterion level of confidence. However, the notion that a wide variety of triples is important in the DG paradigm was also undermined in Experiments 5 and 6 which showed that triple heterogeneity was of itself not related to task success.

How, then, can Evans’ (2006, 2007) hypothetical thinking framework account for DG facilitation? The basic premise that DG instructions overcome the singularity principle remains intact. The singularity principle claims that people can only think about one possibility or mental model at a time, due to hypothetical thinking requiring the use of the analytic system which is limited in capacity and sequential in nature (Evans, 2007). Evans emphasises that it is the combination of a single hypothesis being considered at one time along with the satisficing principle (i.e., that the hypothesis is “good enough”) which leads to poor performance. According to the hypothetical thinking framework, the fact that the initial sufficient hypothesis induced by the exemplar satisfies the requirement of the task means that alternatives do not need to be brought to mind and tested (a view similar in its functional consequences to Friedrich’s, 1993, PEDMIN analysis of the task). The E-ICM presented in this thesis suggests that testing of the DAX and MED hypothesis is sequential; that participants are testing either the DAX or the MED hypothesis and not both. It could be argued that evidence from a MED test providing diagnostic information as to the boundaries of the DAX rule is, in fact, an unforeseen consequence for the participant that was not anticipated when they designed their MED test. Thus although information from a MED test can impact on knowledge of the DAX rule, the participant is at any stage still only testing one hypothesis at a time. According to this analysis DG facilitation can be accommodated within the hypothetical thinking framework, as, indeed, can the E-ICM.

10.8 Future Directions

This thesis has been concerned with exploring DG facilitation within Wason’s (1960) 2-4-6 task. It began by systematically testing contemporary accounts of the effect with a view to discriminating between them. The results of these studies suggested that no single account could provide a wholly adequate account of the effect. As a result of this work a new model was drawn up which extended Oaksford
and Chater's (1994) ICM to account for the effect of DG instructions. It has been suggested that the E-ICM can be accommodated within Evans' (2006) dual processing account of hypothetical thinking, although these ideas are somewhat tentative. Further work is required to test predictions derived from the E-ICM, the role of contrast class triples in success on the SG variant of 2-4-6 task, and the relationship between the number and range of triples generated and the production of a contrast class triple.

The exploration of the role of contrast class triples on the SG variant of the task is an obvious first step; data presented here regarding contrast class information has largely been pooled across instruction mode and success on the task. It is therefore of importance to check that contrast class information promotes success on the SG version of the task. It would be expected that such an effect would be evident, and also that such triples are not commonly produced. One aim of the thesis would thus be fulfilled; the development of a convincing explanation of DG facilitatory effect would go some way to explaining poor performance on the SG variant of the task. This in turn could lead to the development of manipulations which promote success on the SG version of the task. It is of note that Vallée-Tourangeau and New (1997) produced a scenario-based version of the task which successfully manipulated the types of triples produced with resultant differential performance on the task, a finding which was replicated by Sutherland, Lucas, and Gale (1998) in a probabilistic version of the same task. However, reporting of results did not include analyses of the presence of descending triples in either of these studies, and it is known that scenario-based instructions can affect performance on a variety of reasoning tasks (see e.g., Griggs & Newstead, 1982; Wason & Shapiro, 1971). It would therefore be useful to explore this effect more fully.

A second route of enquiry would be to explore the role of descending triples when constraints imposed by this programme of study are relaxed. For example, with the exception of Experiment 1, all participants were required to produce an exact number of triples. That this may affect naturalistic hypothesis testing strategies has been noted several times in this thesis. As the contrast class theory had not been developed when the triple number restriction was introduced it would be useful to explore this aspect of hypothesis testing behaviour along with any concomitant effect on triple variety. In addition, throughout the experiments reported in this thesis only a single attempt at announcing the rule was permitted. However as Vallée-Tourangeau and Payton (2008) note, this prevents participants developing their hypothesis testing strategies when told that their hypothesis is incorrect. This is an important avenue for further exploration as it has been argued that it is likely that strategies for hypothesis
testing may vary as a session progresses (e.g., 2000; Roberts, 1993; Siegler & Chen, 2002).

In the literature review it was noted that two aspects of hypothesis testing behaviour have been identified: the context of discovery and the context of justification (Reichenbach, 1938). Although this thesis has been concerned with both aspects of behaviour, no attempt has been made to discriminate between the two aspects; indeed, hypothesis testing behaviour has been presented as a cyclical activity in which the two aspects are intimately entwined. Contrast class theory, however, is mostly concerned with the context of discovery; it is suggested that it is the production of a contrast class triple which leads to the production of the correct rule. This leads to the intriguing possibility that it might be possible to disentangle these two aspects of hypothesis testing behaviour. One possibility worthy of consideration would be the use of an eye tracking methodology to assess what information is used in hypothesis formation. For example, it would be possible to present participants with protocols from imaginary participants and then use eye tracking data to infer what data was used in generating hypotheses.

In terms of exploring the E-ICM more fully it has been suggested that verbal protocols would be one possible methodology which might be usefully employed to elucidate the processes involved in hypothesis testing and how they relate to E-ICM. Although potential problems with the approach have been highlighted, there remain alternative ways to enhance the methodology and diminish its intrusiveness, for example, by collecting participants' working hypotheses and their reasons for test choices.

To generalise the findings presented in this thesis it would be useful to apply contrast class theory to other hypothesis testing domains outside of the 2-4-6 task. Gorman et al.’s. (1984) New Eleusis task represents an example of a paradigm that could facilitate such generalisation, although there would need to be careful consideration as to what would constitute a contrast class test in this more complex rule discovery context.

One further theme which has emerged from the thesis is that strategies which participants employ may vary as they progress through the task, a position in line with both the ICM and E-ICM which suggest that participants move from using counterfactual testing to generate hypotheses to positive testing of generated hypotheses. Such a notion is also in line with recent work in other reasoning domains such as those involving deductive inference (e.g., Roberts, 1993; Siegler & Chen, 2002). This is an area which could bear fruitful examination. Vartanian, Martindale, and Kwiatkowski (2003b) showed that creativity was associated with
success on the 2-4-6 task, with more creative participants being differentially more successful than less creative participants. Although it is not clear by what mechanism creativity fostered success on the task such findings suggest that differing strategies may impact on success. Perhaps more creative participants are able to move between strategies more effortlessly and it is this aspect of their creativity which underpins success on the task. It is of note that in their analyses Vartanian et al. showed that their participants tended to move from a confirmatory strategy to a disconfirmatory strategy, which is again further evidence that participants are able to vary their hypothesis testing strategies as they progress through the task.

Related to the idea that strategies may change within a hypothesis testing session is the idea that there may be differences in strategies across participants. Indeed informal observation has led to the tentative conclusion that there are at least two distinct approaches to the task: one of exploring the hypothesis space (most commonly employed) and a less common approach of exploring the experiment space (cf. Klahr & Dunbar, 1988). These approaches are manifest by either a tendency to produce triples slowly, interspersed by periods of thought (hypothesis exploration), or by producing a series of triples fairly quickly followed by a much longer period of thought (exploring the experiment space).

These latter observations fit with the work of Klahr and Dunbar (1988), who used a more naturalistic task which involved discovering the purpose of a button on a remote control device they called “Big Trak”. Klahr and Dunbar characterised their participants as either “experimenters” or “theorists” depending on the approach taken to the task. All participants started off by deploying similar strategies, that is, using hypotheses to guide search in the experiment space. They differed, however, in the way that new hypotheses were developed once the initial hypothesis was abandoned. Experimenters produced new trials to allow a search for regularities in experimental outcomes while theorists searched the hypothesis space for a new hypothesis. Klahr and Dunbar suggested that one key determinant of the approach taken by their participants was prior knowledge; they found that all of the “theorists” but only one “experimenter” in their Study 1 had previous programming knowledge which could be drawn on in developing hypotheses. This suggests that a level of expertise may underpin the strategies employed across participants in their task. However, it is unlikely that expertise would underpin any apparent differences in strategy using the 2-4-6 task as it is unlikely that there would be differential levels of expertise in such a general number rule discovery task. It would be interesting, therefore, first to explore if apparent differences are evident across participants on the task (which would be predicted both from findings in other problem solving tasks and from informal
observation of participants), and what particular participant characteristics underpin the use of differing strategies.

Several times within the thesis it has been hinted that one factor which may affect a participant's performance on the 2-4-6 task is working memory capacity. For example, it has been argued that indirect evidence for working memory constraints affecting performance is provided by the work of Vallée-Tourangeau and colleagues on the role of external representation in the task (e.g., Vallée-Tourangeau & Payton, 2008; Vallée-Tourangeau & Penney, 2005). Vallée-Tourangeau and colleagues showed that when some of the mental work involved in the 2-4-6 was externalised success rates in the SG variant rose. One possible criticism of Vallée-Tourangeau and his collaborators' work is that the external representation used also constrained the hypothesis and experiment space as the external representations prevented numbers over 18 being used in triple generation (in one study; other studies had smaller maximum numbers). However, as Vallée-Tourangeau and Payton note, even with this reduced triple-space there are still over 6,000 possible triples in their graphical version of the task, and success rates in the control group which did not have access to external representations were similar to those reported using standard instructions. Studies which systematically examine the effect of working memory capacity on the task may well be a useful avenue for future work.

10.9 Conclusion

This thesis has explored contemporary theories of DG facilitation (e.g., Tweney, Doherty, Womer, Pliske, Mynatt, R.Gross et al., 1980) in Wason's (1960) rule discovery task with the view that a satisfactory explanation for the DG effect would go some way to explaining poor performance in the standard paradigm. Within the thesis a new account of the DG facilitation effect – the contrast class theory – has been developed which has been situated within an extended version of Oaksford and Chater's (1994) ICM. More broadly it has been argued that an explanation of the DG facilitation may lie within a restructuring of the problem which promotes the use of a simplified form of Bayesian testing of competing hypotheses. Direct manipulations have undermined both positivity bias theory (Evans, 1989) and goal complementarity theory (Wharton et al., 1993), and it has been shown that although there is some evidence for both information quantity theory (Wharton et al., 1993) and triple heterogeneity theory (Vallée-Tourangeau et al., 1995), neither provides a wholly adequate account of the effect. These theories can arguably be subsumed within the contrast class theory as developed within this thesis, although it has been
acknowledged that further work needs to be done to explore the viability of this approach.

Along with suggestions for testing the E-ICM developed in this thesis, other fruitful avenues for exploration have been suggested which aim to examine the contrast class theory within the 2-4-6 task as well as in other domains. More generally, ideas relating to individual differences in strategy use on the task have been presented. It is clear that a great deal more research can still usefully be carried out in order to explore hypothesis testing behaviour within the laboratory context.
Chapter 11
References


Mitroff, l. (1974). The subjective side of science Amsterdam: Elsevier


behaviour in a conceptual task. *European Journal of Cognitive Psychology, 6*,
149-169.

constituents. *Journal of Experimental Psychology: Learning, Memory and


Experimental Psychology: Section A, 49*(2), 447-462.


Reich, S. S., & Ruth, P. (1982). Wason's selection task: Verification, falsification and

Reichenbach, H. (1938). *Experience and prediction: an Analysis of the foundations


Roberts, M. J. (1993). Human reasoning: Deduction rules or mental models, or both?

Rossi, S., Caverni, J. P., & Girotto, V. (2001). Hypothesis testing in a rule discovery
problem: When a focused procedure is effective. *Quarterly Journal of

memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36-
71.


References


