Rules and Similarity in the Development of Category Learning in Children and Infants

By
Caspar Addyman

A THESIS SUBMITTED FOR THE DEGREE OF
Doctor of Philosophy

School of Psychology,
Birkbeck, University of London.

May 2009
Originality Statement

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at the University of London or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at University of London or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project’s design and conception or in style, presentation and linguistic expression is acknowledged.'

Signed: ___________________________________________
Abstract

Most theorists agree that adults can learn and use categories that are both rule- and similarity-based. There is less agreement about the mechanisms involved or on the distinction between rules and similarity. We begin by reviewing two recent theoretical positions, that of Pothos (2005) and of Ashby & colleagues (Ashby et al., 1998; Ashby and O’Brien, 2005). Identifying the notion of a rule to be the most problematic concept, we continue with a broader survey of rules in cognition.

We propose that a developmental perspective could provide insight into these problems. We review the existing literature from infancy and early childhood to show that previous studies have only tangentially investigated rules and similarity. We adapt and extend existing methods and present novel procedures and analyses that allow us to focus on a range of questions; When do rules emerge? What form do early rules take? Can statistical mechanisms give rise to rule-like performance? Is there a competition between rule and similarity processes?

Four groups of experiments explore these questions. One set of experiments looks at 5-month-olds’ (and adults’) performance on several artificial grammar tasks developed from Kirkham et al. (2002). We demonstrate that participants are sensitive to the complexity of certain sequence structures. Another pair of tasks investigate infants’ learning of abstract Same-Different relations. We find that 8-month-olds will learn the relation in both an habituation and an anticipation task, whereas 4-month-olds only succeed on the latter. A third group of tasks juxtapose unidimensional and family resemblance categorization using both rich and impoverished stimuli. We find that infants form flexible classifications. Finally, we present a set of computer games for young children that explore these questions in a more ecologically valid setting than traditional laboratory classification tasks.
Contents

Abstract 3

List of Figures 7

List of Tables 12

Acknowledgments 14

Chapter 1. Rules and Similarity in the Development of Category Learning in Children and Infants 15

1.1. Aims and Objectives 17

1.2. Two Views of Rules and Similarity within Category Learning 21

1.3. Pothos (2005) 26

1.4. In cognition, what are the rules? 35

1.5. Rules in Reasoning 41

1.6. Rules in Language 58

1.7. Rules and the Prefrontal Cortex 67

1.8. Rules in Concept Formation 71

1.9. Rules in Animal Cognition 76

1.10. Statistical Criteria for Rule Use 78

1.11. Some Rules about Rules 81

1.12. Some problems with Similarity-based Categories 84

1.13. The Generalized Context Model (GCM) 88

1.14. Why Study Category Learning in Early Infancy? 90

1.15. Methods for Investigating Categorisation in Infancy 92

1.16. Categorization in Infancy - Experimental Findings and Developmental Theories 99

1.17. Rules and Similarity in early childhood and beyond 110

1.18. Conclusions and questions for Research 114
Chapter 2. The role of stimulus complexity in visual sequence learning by five month old infants and adult controls.

2.1. Introduction 118

2.2. Experiment 1 - Random vs. Pair-based Sequences 135

2.3. Experiment 2 - Pairs vs. Triplets, High Overlap 145

2.4. Experiment 3 - Pairs vs. Triplets, Low Overlap 150

2.5. Experiment 4 - Adult Control Task 154

2.6. General Discussion 162

Chapter 3. Infant learning of abstract Same/Different relations.

3.1. Introduction 168

3.2. Experiment 1a - S/D Discrimination in 8-month-olds using Photographic Stimuli in a Habituation Task 176

3.3. Experiment 1b - S/D Learning and Generalization in 8-month-olds using an Anticipatory Looking Paradigm 183

3.4. Experiment 2a - S/D Discrimination in 4-month-olds using Photographic Stimuli in a Habituation Task 196

3.5. Experiment 2b - S/D Learning and Generalization in 4-month-olds using an Anticipatory Looking Paradigm 198

3.6. General Discussion 203

Chapter 4. Unidimensional versus family resemblance category structure in infancy

4.1. Introduction 213

4.2. Experiment 1 - 11-month-olds with artificial stimuli from Smith (1981) 216

4.3. Experiment 2a - 11-month-olds with Photographic Stimuli 223

4.4. Experiment 2b - 5-month-olds with Photographic Stimuli 229

4.5. General discussion 233


5.1. Introduction 237

5.2. Ecological Validity in the study of Category Learning 238

5.3. Technical Background 250
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4. Experimental Investigation - Developmental Differences in the</td>
<td>256</td>
</tr>
<tr>
<td>Learning of Rule And Similarity Category Structures</td>
<td></td>
</tr>
<tr>
<td>5.5. Experiment 1 - Alien Grab - Category learning with feedback.</td>
<td>263</td>
</tr>
<tr>
<td>5.6. Experiment 2 - Alien Sort - Free category construction with prior</td>
<td>270</td>
</tr>
<tr>
<td>familiarisation</td>
<td></td>
</tr>
<tr>
<td>5.7. Experiment 3 - Manual free-sorting task</td>
<td>280</td>
</tr>
<tr>
<td>5.8. General Discussion</td>
<td>283</td>
</tr>
<tr>
<td>5.9. Future Directions</td>
<td>286</td>
</tr>
<tr>
<td>Chapter 6. Conclusions</td>
<td></td>
</tr>
<tr>
<td>6.1. Summary of Findings</td>
<td>290</td>
</tr>
<tr>
<td>6.2. Synthesis</td>
<td>295</td>
</tr>
<tr>
<td>6.3. Questions for Future Research</td>
<td>299</td>
</tr>
<tr>
<td>6.4. Lessons Learned</td>
<td>301</td>
</tr>
<tr>
<td>6.5. Final Remarks</td>
<td>302</td>
</tr>
<tr>
<td>Bibliography</td>
<td>305</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Examples of stimuli from information integration and prototype distortion categorization tasks. 23

1.2 Examples of explicit and implicit categorization from Waldron and Ashby (2001). 24

1.3 Schematic representation of the two phase model of rule-based reasoning 39

1.4 Schematic representation of the modular architecture of ACT-R (version 5.0) with an indication in brackets of the brain areas these are thought to correspond to. From Anderson et al. (2004) 49

1.5 Two models of English past tense formation. See section 1.6.1 for details. Diagrams taken from McClelland and Patterson (2002b) and Pinker and Ullman (2002b) respectively. 62

1.6 Rules in the PFC. Diagram from Bunge and Zelazo (2006). 69

1.7 The relationship between rules and similarity in categorization as envisaged by Hahn and Chater (1998). 75

1.8 Four possible patterns of rule learning. (a) The rule boundary is discrete and known from the beginning, learning involves reduction in errors. (b) Rule is known from beginning, learning determines the boundary more precisely. (c) Learning is case by case, resulting in rule-like pattern in the limit. (d) Rule boundary and errors are both gradually improved with learning. 84

1.9 Examples of prototypes and levels of distortion used in Younger and Gotlieb (1988). 100

1.10 Some stimuli from Younger (1985). 103

1.11 A range of possible triad types from Thompson (1994), experiment 1. 113

2.1 The relation between successive shapes in Kirkham et al. (2002). 127
2.2 Infants’ Looking Time to Familiar and Novel patterns. Kirkham et al. (2002) 128
2.3 The increasing entropy of four sequence types with sequence length. 131
2.4 Comparing the two redundancy measures for a window of six items. 134
2.5 Graph of overall results for Experiment 1. Mean looking times (in seconds) for random and structured colour and shape sequences for 5 month old infants. There was a decrease in looking over the three pairs of trials and infants consistently looked longer at the random sequence. Error bars are 95% confidence intervals. 138
2.6 Graph of results for Experiment 1 grouped by gender. 139
2.7 Graph of results for Experiment 1 grouped by a median split on total looking time. 140
2.8 Graph of the zero order redundancy scores calculated using Equation 2.15 with a window size 8. See text for explanation. 142
2.9 Graph of the relative frequency scores for single and pairwise items. See text for explanation. 143
2.10 Graph of mean looking times in Experiment 2 to pair and triplet based sequences grouped by gender. 147
2.11 Graph of mean looking times in Experiment 2 to pair and triplet based sequences grouped by sequence order with trials in exact order they were shown. 147
2.12 Example of the type of sequences used in Experiment 2. Note that a procedural error lead to the two types of sequence having maximal grammatical overlap. 148
2.13 Graph of mean looking times in Experiment 3 to pair and triplet based sequences grouped by sequence order with trials in exact order they were shown. 151
2.14 Graph of the relative frequency scores for single and pairwise items in Experiment 3. See text for explanation. 153
2.15 Illustration of the transition probabilities in the four types of test sequence. See text for explanation. 156
2.16 Bar chart of the accuracy of participants in the adult sequences task. Each column represents the mean correct responses for each of the possible sequence pairs, grouped by gender. 158

2.17 The relative proportion of correct responses in Experiment 4, grouped by participants confidence in their answer. 160

3.1 The picture stimuli used in Experiment 1a. 178

3.2 Procedure for Experiment 1. 180

3.3 Results from the Same/Different habituation Experiment 1a. 181

3.4 A schematic representation of the anticipatory looking experiment. 187

3.5 An example of all the fixation data collected from a single participant. 189

3.6 Summary of the proportion scores for all participants in Experiment 1b averaged by block. 190

3.7 Relative proportion of available fixation data where infants were looking to the correct ROI 192

3.8 Proportion of first looks that were to the correct ROI 193

3.9 Graph showing the different accuracy for during learning for the different pairs of stimuli for 8 month old infants in Experiment 1b. 194

3.10 Graph showing the different accuracy for trials where the pairs of objects were either the same or different from each other. The first two bars show the overall accuracy Error bars show the 95% confidence intervals. 195

3.11 Results from the Same/Different habituation experiment with 4 month old participants. 197

3.12 Summary of the proportion scores for all 4 month old participants in Experiment 2b averaged by block. 200

3.13 Relative proportion of available fixation data where 4 month old infants were looking to the correct ROI, averaged across all included participants and grouped by blocks of four trials. Error bars show then 95% confidence intervals. 200

3.14 Graph showing the different accuracy for during learning for the different pairs of stimuli for 4 month old infants in Experiment 2b. 201

3.15 Graph showing the different accuracy for trials for four month old infants 202
4.1 Example stimuli based on Smith (1981) showing five steps of variation in each of four dimensions (height, roof pitch, arrow orientation & colour) 217

4.2 Binomial frequency distribution for each variable dimension in Experiment 1a. Given dimensional values -2, -1, 0, 1 & 2, the features could take these [1, 4, 6, 4, 1] times respectively. 218

4.3 Illustration of how test stimuli are determined online for each participant. 219

4.4 Design for Experiment 1a. 219

4.5 Mean looking times in seconds for Experiment 1. There was a highly significant difference between average of the first three and last three habituation trials indicating that habituation had occurred. However, at test the infants only dishabituated to the novel test item. 221

4.6 Design for Experiment 2a. 225

4.7 Mean looking times in seconds for Experiment 2a. There was no significant difference between average of the first three and last three habituation trials indicating that habituation did not take place. At test, infants looked longer at the out of category test item and the novel item. 226

4.8 Results of Experiment 2A grouped by type of familiarisation trial. On the left, mean looking times for first and last three habituation trails. (Shown for comparison to test trials.) On the right, the interaction of habit set (cars vs houses) with test trail (In category, Out category, Totally Novel). 227

4.9 Results of Experiment 2B. On the left, mean looking times for first and last three habituation trails. On the right, mean looking times for the three different test trial types. Also shown are 95% confidence intervals and the results of one tailed t-tests. See test for explanation. 230

4.10 Graph showing the combined results of Experiments 2A and 2B 231

5.1 Two exemplars illustrating the 9 variable dimensions of the alien stimuli. 262

5.2 Screenshot from the Alien Grab game. 264

5.3 Summary of results from Experiment 1. The graphs show the mean accuracy per round for each age group for the three different types of category structure; (a) similarity, (b)1D rules, and (c) 3D rule. 266
5.4 Summary of results from Experiment 1. The graphs plot the mean number of aliens collected per round for all participants grouped by the category structure they saw. Figure (a) shows the results for the 2 similarity conditions and Figure (b) shows the four rule-based conditions.

5.5 Graph showing the mean rate of collection of aliens per round grouped by age in Experiment 1.

5.6 Screenshots from the Alien Grab game used in Experiment 2.

5.7 Schematic depiction of the predictiveness scores were calculated for the free sorting task in Experiment 2.

5.8 Mean dimensional predictiveness scores for the six category structures sorted by all game participants in Experiment 1B. Error bars are 95% Confidence intervals.

5.9 Comparing the mean dimensional predictiveness scores in game (light blue) and manual (red) sorting tasks by Year 6 participants. There were insufficient data for the other comparisons. Error bars are 95% Confidence intervals.
List of Tables

2.1 Redundancy and relative frequency measures for Experiment 1. See text for explanation. 142
2.2 Redundancy and relative frequency measures for Experiment 3. See text for explanation. 152
2.3 The differences in entropy for the 4 sequence types compared in Experiment 4. 158
2.4 Results of a Logistic Regression Analysis of adults’ ability to discriminate two sequence types. Shown are the parameters for a model based on a measure of difference in entropy between sequences and a parameter for the gender of the participants. 159
3.1 Significance scores by participant by block for Experiment 1b. All values are for two-tailed, one-sample t-tested comparing mean proportion scores for each individual to the chance value of 0.5. Significant values are shown in bold. 190
5.1 Feature structure used in Medin et al. (1987) and numerous other category learning experiments. A family resemblance sort is possible, as shown, grouping the two prototypes A0 and B0 with the ‘one-aways’, items which are different from prototype in just one feature. Notice that uni-dimensional partitions on any one of the four dimensions are also possible. 239
5.2 Comparison of the typical features of category learning as encountered in the laboratory and in the real world. 244
5.3 The six category structures used for the aliens in Experiments 1 and 2. 261
5.4 Table showing distribution of the most predictive dimension for each participant across category structure conditions. The final column indicates the total number of participants in each condition. 275
5.5 The distribution of uni-dimensional and multidimensional sorts shown for various levels of cut-off the relative importance of the second dimension. See text for fuller explanation. 277

5.6 Comparing the distribution of uni-dimensional and multidimensional sorts shown for various levels of cut-off the relative importance of the second dimension for the Year 6 participants in Experiments 2 and 3. 281
Acknowledgments

The greatest thanks are due to my supervisor, Professor Denis Mareschal, who has guided me through every stage of this research with support, encouragement and even constructive criticism.

I am very grateful to the parents and infants who participated in testing at the Birkbeck Babylab and to the adult volunteers and the staff and pupils of the schools that took part in other studies. I also acknowledge the immense benefit this research has received from the efforts of Leslie Tucker and the support staff of the Babylab.

Likewise, I owe a debt to my many friends and colleagues at the Centre for Brain and Cognitive Development and to its director Professor Mark Johnson for making it such a fun and stimulating place to work. Dick Aslin and Ulrike Hahn provided valuable discussions at various stages of the project. The following people provided particular help or assistance either with testing, coding and in other more intangible ways; Jennifer Betters-Bubon, Fani Deligianni, Sarah Lloyd Fox, Sarah Cayless, Julia Kennedy, Nick Boyes, Jennie Brown, Kathrin Cohen-Kadosh, Belinda Brown & Tessa Dekker.

This research was completed as part of the European Commission funded collaboration (the FAR project) and has benefited greatly from the collaboration and interchange of ideas with other members of the network. It was made possible by EC Framework 6 NEST contract 516542.

Finally, I thank my family for making my own childhood such a happy one and continuing to support me in everything I do.
CHAPTER 1

Rules and Similarity in the Development of Category Learning in Children and Infants

_The Purist_

I give you now Professor Twist,
A conscientious scientist,
Trustees exclaimed, "He never bungles!"
And sent him off to distant jungles.
Camped on a tropic riverside,
One day he missed his loving bride.
She had, the guide informed him later,
Been eaten by an alligator.
Professor Twist could not but smile.
"You mean," he said, "a crocodile."

Ogden Nash, _I'm a Stranger Here Myself_ (1938)

Ogden Nash never informs us what his conscientious scientist was sent into the jungle to do. But we may guess, based on our familiarity with other jungle-bound scientists, that Professor Twist was probably a biologist, sent there to classify the fauna he found. Classification can be done in two ways. One may place like with like, as when we see the similarity between a crocodile and an alligator. Or else one may follow rules, as when a biological taxonomist draws up the list of the defining characteristics of _Crocodylus niloticus_ and _Alligator mississippiensis_. It might seem obvious that these are two very different processes; The first is fast, flexible and automatic while the second is precise, focused and deliberate. But how do they work, how are they related and how do we determine which type of categorisation to use and when?
Classification by means of similarity seems like child’s play, but this belies its power. It is the work of an instant to identify to an alligator, which is just as well because it might be a matter of life or death. We could still recognise a crocodile if it was bright pink, or half an inch long, a character in a cartoon or even if it had been turned into a handbag. Even young children can do this and all without being able to define or articulate what is the essence of a crocodile, or even needing to. Similarity allows one to generalise. If you had only ever met alligators you would have no difficulty in recognising that crocodiles were their cousins and if you were told that alligators have integumentary sense organs, you would be likely to conclude that crocodiles did so too. One may mistake a crocodile for an alligator but one would not make the mistake of thinking that either was friendly. Although sometimes, like perhaps the unfortunate Mrs. Twist, one may be fatally misled by their shared similarity to floating tree trunks.

And what of classification by rules? Well, take for example, alligators and crocodiles; the similarity of the two is such that it takes a rule to tell them apart. There are a few rules you could choose from; Alligators have U-shaped snouts while crocodiles have pointy V-shaped noses. Crocodiles have teeth in their lower jaws that protrude up the sides of their mouths while alligators do not. Or, since Professor Twist did not see the smiling face of the beast that ate his bride, we may reason that he reasoned by another rule; “If one is in Africa then they are crocodiles.” All involve the application of abstract knowledge to a restricted set of the available information. The first two involve focusing on a small number of perceptual features and the third uses a verbal rule that concentrates entirely on context. While classification rules may then appear the remit of quintessentially rational adults, they are not purely for pedants. By building categories that are based on single simple features, or definitions, we acquire manageable, manipulable blocks with which to reason. Rules can be chained, nested, combined or even contradicted to build up complex knowledge about the world. By these means, conceptual thought is extended beyond its perceptual foundations in a way that is distinctively human.

Like Professor Twist, the present author is also setting out conscientiously on a mission of scientific classification. In fact the aim of this thesis is a classification rule.
of classifications. It sets out to tell the difference between rules and similarities in category learning, to understand how each process works and how they are related. A biologist seeking to understand alligators and crocodiles can investigate their anatomy, physiology, life-cycle and ecology. He or she can also appeal to the theory of evolution and look back into the fossil record. Likewise, the cognitive scientist can meticulously investigate fully realised adult abilities but may also appeal to comprehensive theories of these abilities and look at the early origins of adult-like performance in the abilities of children and infants. This thesis adopts that developmental approach. The rest of this chapter provides an introduction to the relevant theoretical and empirical literature and motivates the particular developmental approach taken. The following four chapters each presents a self-contained piece of empirical research that aims to address some of the questions raised. While the final chapter summarizes the results that are found and considers what more general conclusions can be drawn. The rest of this section lays out in more detail the particular the aims and objectives of this research and provides a brief outline of the literature review that follows.

1.1. Aims and Objectives

The primary objective of this thesis will be an empirical investigation of the development of rules and similarity in category learning in children and infants. Categorisation is the ability to treat discriminably different objects in an equivalent fashion, to group items according to common features (Murphy, 2002). Previous research has shown that in adults there are clear distinctions to be found between categorisation by similarity or by rules. In the former case groupings are made according to associative mechanisms that take into account the overall similarity of items based on many of their features (e.g. the category of houses, of dogs, cats or chairs), while in the latter case groupings may be based on explicit narrow criteria or single features (e.g. the category of all red objects, objects larger than a house, or animals with four legs). Alternatively, rule-use may be identified by evidence of abstract symbol manipulation, as in the case of language. In children, the boundaries between rules and similarity are less clear and in infants it is unclear if rules are used at all.
Therefore, this research aims to investigate the earliest emergence of rule use in infancy and the development of dual rule- and similarity-based categorisation mechanisms in early childhood. The objective will be to approach the general problem from three specific directions. Firstly, by examining pre-cursors to explicit rule use in pre-linguistic infants and, in particular, by looking at infants' abilities to learn and generalise very simple rules. Secondly, by comparing infants' performance on categorisation tasks that require single or multiple dimensional category groupings, looking for evidence that infants are able to selectively utilize rule-based or similarity-based classifications. Thirdly, by conducting studies which directly compare how young children and adult participants respond to rule- and similarity-based categories.

Furthermore, the emphasis will be on understanding infant categorisation abilities from a developmental perspective. It is not sufficient simply to catalogue the abilities of different ages groups, an attempt must be made provide a theoretical explanation for the pattern of results discovered. This is important both in terms of understanding the changes occurring during infancy and when considering how the infants’ abilities are to be related to fully adult competencies. Therefore, the proposed experimental investigations in addition to looking for the earliest emergence of rule-like behaviour and of dual rule and similarity mechanisms. It will also look for developmental change in the first year of life and attempt to relate this development to what is known about eventual adult abilities.

However, investigating infant abilities necessarily requires special techniques appropriate to the highly restricted cognitive abilities of infants. All theories must be assessed in the face of the many methodological limitations inherent in the testing of infants. Hence, attempts to provide continuity between infant and adult findings are highly constrained by these limitations. It is very unlikely that theories applicable to adult category learning abilities can be tested directly in infants and it will be equally rare that one can develop an equivalent adult control for any given infant task. This need not be a major cause for concern since, to borrow the unfortunately loaded terms of Chomsky (1965), infant *competence* will have implications for adult *competence* while no analogous claim need hold for infant and adult *performance*. If infants can be shown to understand some rule, this is of primary theoretical importance, it is of secondary importance whether this is
the same mechanism as is ultimately in operation in adults. In fact, the expectation is that thanks to experience, automatization and representational redescriptions (Karmiloff-Smith, 1992), adult mechanisms will be different from those of infants. But evidence of infant competence can contribute to an explanation of how those adult abilities arise. Nevertheless, the gap is still large and, therefore, a final strand of research will look at the abilities of young school age children (between 5 and 12 years of age) and compare them directly to adults on a common set of tasks.

1.1.1. From Associations to Rules in the Development of Concepts (FAR). This research is part of a larger EC-funded collaboration that aims for a broad understanding the emergence of rule-based behaviours from associative similarity judgments. The 'From Associations to Rules in the Development of Concepts (FAR)' project comprises a network of researchers at six European universities with wide range of expertise all interested in how rule-based systems develop. The Birkbeck group is interested in developmental aspects of the question while other groups are studying category learning in adults (Cardiff and Swansea) and animals (Exeter), and looking to model the emergence of rule use with statistical (Amsterdam) and connectionist (Burgundy) techniques.

The project starts from the notion that rule use and abstract symbol manipulation are typically portrayed as abilities unique to adult humans whilst non-linguistic species and pre-linguist humans appear to operate by means of associative or similarity-based processes. Nevertheless, rules must have their origins in earlier abilities and the overall aim of FAR project is to understand these origins both developmentally and phylogenetically. The developmental work presented here will be highly informed by the insights of these other areas and, in turn, it is hoped that this research will provide a key element in an integrated understanding of the origins of fully adult rule use. In particular, that an understanding of infant competencies will form a bridge between the associative learning abilities of animals and the unequivocally rule-based abilities of adults.

1.1.2. Chapter Overview. The rest of this chapter will lay the theoretical and methodological groundwork for the empirical research presented in Chapters
2 to 5. It surveys existing literature on rules and similarity and provides an introduction to the investigation of category learning in infancy and childhood. The remainder of the chapter can be divided into roughly five parts.

(1) The debate about rules and similarity. First, we survey two theories that aim to account for the contrast between rules and similarity. In Section 1.2 the neurobiological-based theory of Greg Ashby and colleagues (Ashby et al., 1998; Ashby and Ell, 2001; Waldron and Ashby, 2001; Ashby and Spiering, 2004; Ashby and O’Brien, 2005) is introduced which associates rules and similarity with different brain areas. This is contrasted with the continuum hypothesis of Emmanuel Pothos (Pothos, 2005), described in Section 1.3, which sees rules and similarity as the opposite ends of continuum. Rules being associated with the use of a small subset of available psychological dimensions while similarity requires broadly equal weighting of all available features.

(2) Survey of Rules. Sections 1.4 through 1.11 present a fairly extensive survey of rules in cognition. By looking at the notion of rules across a range of cognitive domains, it is hoped to establish a clearer picture of what exactly rules are and how they are used. This exercise in the clarification of the concept of a rule is particularly important prior to an investigation of rules in infancy, where the 'intuitions' that apply to adult rule use are no longer so obvious. We consider historically important conceptions of rules within cognitive science and look at how rules have been defined across a range of disciplines with a view to how these perspectives may inform our developmental investigation of early pre-cursors to adult rule use.

(3) Similarity. Next follows a shorter survey of similarity. Section 1.12 looks at some problems with the notion of similarity and its relation to category learning. While Section 1.13 gives a short overview of the Generalized Context Model (GCM, Nosofsky, 1986), a highly successful exemplar-based (i.e. similarity-based) model of categorization.

(4) Category learning in infancy. In Sections 1.14 through 1.16, we consider why it is appropriate to investigate rules and similarity from a developmental perspective and look in detail at existing research on category learning in infancy. This includes a survey of the methods available to conduct this type of research and a
review of the findings and theories within this field. It shows that rules and similarity have not been investigated directly within the field of infant category learning but that certain results suggest it would be possible and fruitful to do so.

(5) Rules and similarity in early childhood. Finally Section 1.17 looks at existing research on rules and similarity in early childhood. It highlights some of the limitations of existing methods and makes the case for research that directly compares the performance of children and adults on rule- and similarity-based tasks. While Section 1.18 draws some general conclusions and focuses on the key questions that ought to be considered in the empirical work that follows. Note that each experimental chapter will also feature a substantial survey of the literature specifically relevant to the particular empirical questions it addresses.

### 1.2. Two Views of Rules and Similarity within Category Learning

In a series of papers, Ashby and colleagues (Ashby et al., 1998; Ashby and Ell, 2001; Waldron and Ashby, 2001; Ashby and Spiering, 2004; Ashby and O’Brien, 2005) present experimental, neuropsychological and neurobiological evidence for category learning being mediated by the competition between distinct rule-like and similarity-like mechanisms. They restrict their attention to the learning and acquisition of new categories and do not address expert performance or rules and similarity distinction in other domains. By contrast, a differing theoretical perspective was set out by Pothos (2005). He sees rule-like and similarity-like behaviours as opposite extremes on a continuum. Rule-like performance is claimed to be associated with decisions based on the salience of just one or only a few properties of the stimulus, whereas similarity judgements integrate information across a wider number of dimensions. Pothos reviews potential evidence for this continuum across learning, reasoning and language, as well as within the domain of categorization. This review will focus primarily on the application of this theory to categorization.

There is a potential problem with both these theoretical approaches in that the definitions of both rules and similarity become functional. Namely, the distinction is made based on what type of processing appears to be taking place rather than any prior definition of what would constitute a rule or a similarity based task or performance, (for further discussion see Hahn, 2005). To some extent, this
failure to strictly categorize behaviours as either rule-based or similarity-based is the deliberate intent of Pothos (2005), for whom there will always be a wide range of intermediate behaviours. Likewise, although Ashby et al. are clear about existence of separate rule and similarity mechanisms in the brain, a crucial feature of his model is that these are in competition, so behaviour will often reflect both styles of processing. Nevertheless, both agree that there are some qualitatively different observable behaviours and that there exist canonically rule-like and similarity-like tasks.

1.2.1. The neuropsychological model of Ashby and colleagues. Ashby and O’Brien (2005) classify four different types of mechanism; (1) explicit processes for rule-based categorization that use executive attention and working memory, implicit information-integration processes with (2) a small or (3) a large numbers of exemplars and finally (4) prototype distortion mechanisms. Firstly, there are rule-based tasks which generally rely on individuals paying selective attention to no more than a few stimulus dimensions independently. For example, deciding that a figure is a red square, one must assess its shape and its colour, but can verify each property independently of the other. In information integration tasks, such as for example figure 1.1 (a), it is not possible to pay attention to each dimension individually discrimination must be on the basis of an integration across the dimensions. Ashby and O’Brien (2005) make a further distinction between information integration tasks with small and large numbers of exemplars, in the former case a memorization strategy may be possible whereas in the latter it is not. Finally, there are prototype distortion tasks such as the classic studies of Posner and Keele (1968), illustrated in Figure 1.1 (b). In these cases, exemplars of one category are each systematically distorted versions of an unseen prototype.

Ashby et al. (1998) proposed that category learning on the basis of either rules or similarity are very different processes that make use of several different brain systems and that these multiple systems will compete. Rule-like discriminations will be better handled by verbal or declarative systems and similarity judgements will be more suited to implicit or procedural-learning systems. There will also be a competition mechanism that decides which approach can solve a given problem. They map this model onto three brain areas; the prefrontal cortex (PFC), the
inferotemporal cortex (IT) and systems within the basal ganglia. In later versions of the theory, rule-based tasks are defined as "those in which subjects can learn via some explicit reasoning process" (p.205, Ashby and Ell, 2001). Typically these are tasks where only one or two stimulus dimensions are relevant to categorization. They observe that the majority of tasks used in neuropsychological assessment have been of this type and so differences in deficit across conditions have not generally been noticed.

They propose a formal neuropsychological model of how such a two process competitive system might operate in the brain called COVIS (competition between verbal and implicit systems). They use their model to account for dissociations in cognitive behavioural data and patterns of neuropsychological deficit. Ashby et al. (1998) also provides an explicit computational implementation of the model with which they simulate and test certain category learning results. An important feature of the COVIS model is that both/all systems are always active, conscious verbal systems will dominate initially but the implicit systems are capable of better performance, and so will be favoured with further training.

1.2.2. Cognitive and behavioural data. One distinction between rule and similarity judgements is revealed by the effect of feedback on learning performance. Ahn and Medin (1992) showed that when trial by trial feedback is absent, people tend to categorize stimuli on the basis of just one dimension. If feedback is available but is delayed by five seconds or more, learning is impaired on information integration tasks but not on rule-like tasks (Ashby et al., 1999). However, if feedback is
Some of the strongest behavioural evidence for a two systems model comes from Waldron and Ashby (2001). They compared the performance of individuals learning single and multi-dimensional categorization rules (see figure 1.2). Crucially, they found that a concurrent numerical Stroop task impaired performance on the simple or explicit case far more than it did for the implicit case. Stroop tasks are well known to load working memory and PFC systems. This selective effect on one-dimensional (rule-like) categorization tasks implies that these also require working memory systems whilst similarity tasks do not, in line with Ashby’s multiple systems theory.

Ashby et al. (2003) applied traditional tests for procedural learning to rule and similarity tasks and found that only categorization by similarity uses procedural mechanisms. Participants initially learnt either a rule-based or information-integration
categorization. Some participants then switched hands on the response keys whilst for others the locations of the keys were switched. Neither manipulation impaired the rules task group suggesting they had encoded the category as an abstract label. Whereas changing the position of the keys causes a significant decrease in accuracy for the information-integration group. This encoding of response location is found in other procedural learning tasks.

1.2.3. Neuropsychological and neurobiological data. Using converging evidence from development, neuropsychology, single cell and lesion studies in animals and brain scanning work, Ashby and colleagues map out a model of the brain areas involved in category learning and how they are connected and interact. They hypothesize that three main brain areas involved; the inferotemporal cortex (IT), the basal ganglia and the prefrontal cortex (PFC).

The inferotemporal cortex is a high level area of the visual cortex, the termination of the ventral or ‘what’ pathway. Ashby and Ell (2001) suggest that only prototype-distortion type learning takes place in IT, citing an fMRI study that found evidence of learning related changes in IT during a prototype-distortion task (Reber et al., 1998). Ashby and colleagues believe that although the IT represents visual stimuli, the systems for categorization and discrimination are likely to be elsewhere, most likely in the basal ganglia. Hence, although lesions in IT can result in category-specific agnosias, this does not imply that category learning occurs in the visual cortex. For example, single cell recordings in IT cells of monkeys showed cells selectively firing to two visual stimuli associated with ‘good’ and ‘bad’ rewards. However when the pairing was switched the IT cells continued firing as before, indicating the assignment of semantic category was most likely occurring elsewhere (Rolls et al., 1977). The inferotemporal cortex projects into the tail of the caudate nucleus in the basal ganglia.

Ashby and colleagues identify systems in the basal ganglia as essential to category learning. The tail of the caudate nucleus is said to be involved in similarity judgements and the head of the caudate nucleus, which connects to the PFC, involved in rule-like tasks. They cite various sources of evidence. Scanning studies show these areas to be active during category learning (e.g. Rao et al., 1997). Rats with lesions to the tail of the caudate nucleus but intact visual cortex are unable
to learn a two item visual discrimination task but could learn a spatial discrimination (Packard and McGaugh, 1992). Whereas Eacott and Gaffran (1991) lesioned all pathways from the visual cortex except to the caudate nucleus and found visual discrimination was unaffected. The caudate nucleus also has projections from the auditory and somatosensory cortices. But very little research exists into category learning via these sensory modalities.

The prefrontal cortex is one of the last brain systems to come on-line in development, and thus COVIS predicts that young children will have difficulty in rule-switching and also in rule-like categorization tasks. Some evidence supports this view. Young children will preservate on incorrect rule (e.g. Kemler, 1978) However, although initial studies seemed to suggest that young children group by overall similarity rather than specific stimulus dimension (Smith and Kemler, 1977), reanalysis of early data and more recent results suggest that young children actually fixate on single dimensions and apply rules inconsistently (Thompson, 1994). Nevertheless, from the perspective of the current investigation, the link between PFC and rule use could lead to dissociations in development between the rule- and similarity-based classifications. We return to this question in Section 1.7.

1.3. Pothos (2005)

1.3.1. General. Pothos (2005) takes a far more abstract and theoretical approach than Ashby and colleagues. He is interested in finding a broad conceptual framework that can be applied to rule-like and similarity-like behaviour across numerous domains. He applies the framework to categorization as well as learning, reasoning and language. He assumes that the distinctions between rules and similarity are fundamentally the same in each of these domains but concentrates on an analysis of categorization to illustrate his theory.

The main thesis of Pothos has three elements. Firstly that canonical rule-like and similarity-like processes are both extremes of the same continuum, determined by the number of dimensions that are relevant to the decision process. Secondly, that there will also be behaviours that lie in the middle of this continuum. Thirdly, categorization by rule and similarity do not just act on perceptual inputs but can operate over more abstract properties. Furthermore, there are two important assumptions
1.3. POTHOS (2005) 27

supporting this framework; that "the object properties relevant in deciding how to
categorize it amongst a number of candidate categories are the properties uniquely
common to the instances of each category" (p.2, Pothos, 2005) and that "it is pos-
sible to represent objects in terms of discrete entities (perceptual features, abstract
properties, and so forth)” (p.2, ibid.). Based on these assumptions, the core of the
Pothos thesis is that

“... given a set of categories and an object, we can establish a par-
ticular representation of the object relative to these categories that
could include perceptual properties, abstract properties, or both.
We can then categorize the object as a member of one of the cate-
gories, or not ... We postulate that when the object categorization
is determined by a small subset of the relevant object properties,
then categorization should be understood as a rules process. By
contrast, when categorization is determined by most of the relevant
object properties, broadly equally weighted, then categorization is
best understood as an overall similarity process.”

p.2, ibid.

Pothos then shows how these definitions lead to the different characteristics of cog-
nitive processes that follow a broadly rule-like or similarity-like profile. Following
Sloman and Rips (1998), he takes rules to be associated with certainty, composition-
ality, systematicity and productivity, whilst similarity is characterised by flexibility
of judgement. For example, in a rule-based task where only a single dimension is
relevant it is possible to have greater certainty about a categorization but there
is likely be a restricted set of cases where it is relevant (less flexibility). Likewise,
because rule judgements are simpler, it is possible to combine them (composition-
ality), apply them consistently across cases (systematicity) and apply them to new
situations (productivity).

Pothos considers the possibility that, contrary to his hypothesis, only rule-based
mechanisms may give rise to abstract knowledge whilst associative learning will only
lead to similarity knowledge. Evidence from animal studies appears to show that
associative mechanisms do not give rise to abstract knowledge (e.g. Wasserman and
Miller, 1997). However, Pothos argues that if abstract properties (e.g. relationships
between objects) can be perceived or are otherwise available to a learning mechanism then either rules or similarity judgements are possible. Similarly, a study by Shanks and Darby (1998) suggested that human associative learning is a similarity process. Again, Pothos suggests by including relational properties as well as perceptual properties the mechanism can lead to either rule- or similarity-like performance. This underscores the important feature of the Pothos proposal, namely that the analysis should be in terms of psychological dimensions, how individuals represent objects not how they perceive them.

1.3.2. Rules and Similarity in categorization. When specifically addressing categorization, Pothos reviews the positions of other researchers and discusses how their findings might be reanalysed in terms of his rules and similarity framework. Some data in concept/category formation can be explained by reference to either necessary or sufficient critical features that individuals use to assess category membership in certain circumstances (e.g. Rips, 1989, 2001). But this is very different from the classical view of concepts as definitions in terms of a list of necessary and sufficient features (e.g. Katz, 1972). If \( f \) is a necessary feature of concept \( C \) then an object must have at least \( f \) in order to be a \( C \). If \( f \) is a sufficient feature of \( C \) then knowing an object has \( f \) implies it must be a \( C \). In an earlier paper, Pothos and Hahn (2000) demonstrated that analysis of categorization in terms of necessary and/or sufficient critical features fails to capture the full range strategies and behaviours that individuals use. This is consistent with the current Pothos’ theory where in some circumstances critical features may be equivalent to verbal descriptions of a concept or a rule-like behaviour but cannot always determine the psychological representation of concepts.

Similarity in categorization has been best accounted for in terms of exemplar theory (e.g. the Generalized Context Model, GCM, Nosofsky, 1986, 1991, see Section 1.13 below). Generally, these findings are consistent with Pothos’ theory, in part because of the power and flexibility of exemplar models, which allow similarity to be determined by the combination of many equally weighted properties or by selective weighting of only a couple of dimensions to approximate rule like behaviour. In particular, rules will fit into an exemplar scheme if they are thought to be the equivalent of category boundaries orthogonal to the psychological dimensions of the...
representation as suggested by Erickson and Kruschke (1998). Since, the similarity proposal and the behaviour of exemplar models are the same for wholly similarity judgements then data showing that exemplar models can account for similarity supports Pothos’ theory.

1.3.3. Rules and Similarity in reasoning, language and learning. Pothos (2005) also considered a possible continuum approach to rules and similarity in reasoning, language and general learning. This section briefly reviews several of the critical points in each of these areas.

Within the literature on logical reasoning and decision-making, Pothos concentrates on rules-like reasoning since case based reasoning (CBR) approach to problem solving is already congruent with his similarity proposal and so is not considered in detail. The classical approach to logical reasoning would seem to implicate rules alone but various results suggest that actual human reasoning does not follow the rules of formal logic. The Wason selection task (Wason, 1960) being the classic example where heuristics, confirmation bias and pragmatic considerations appear to influence human reasoning. If one looks at problem solving in terms of the properties relevant to making a decision, then when conclusions are drawn on the basis of a small subset of properties the performance will be rule-like independent of concreteness, familiarity, etc of the problem space. Whilst deciding when a particular condition might apply is seen as a similarity process. In this way, Pothos sees most performance in reasoning tasks as a requiring a mixture of both rule-like and similarity-like abilities.

In respect of learning, Pothos restricts his attention to artificial grammar learning (AGL). Here, participants learn to categorize strings of symbols as grammatical or non-grammatical based on prior exposure to a subset of grammatical examples. Where the ‘grammar’ is a set of rules about the order that the symbols occur in. Reber (1967) suggests that participants are in fact acquiring a partial representation of the rules. Dulany et al. (1984) found that the self-reported rules that individuals claimed they were using could account for their accuracy on the task, although these sets of ‘micro-rules’ were often very limited in scope and validity. However, Vokey and Brooks (1992) found that similarity effects influenced participants judgments and Pothos and Bailey (2000) showed that modeling AGL tasks using the GCM
could explain observed variance in participants selections. In both cases analysis of participant data showed that both rule-like and similarity-like effects were present. Pothos’ conclusion is that similarity operates as co-occurrence statistics are noticed but that the more frequent and salient occurrences will give rise to rules, which in some cases can be highly abstract.

From Chomsky (1957) onwards there has been extensive debate on whether language acquisition requires a set of abstract rules of syntax, grammar and inflection or if these regularities are the result of statistical learning mechanisms (e.g. Rumelhart and McClelland, 1986; Elman et al., 1996; Marcus, 2001 ). However, these issues are largely beyond the scope of the Pothos review. He makes the case that any successful connectionist model of any language phenomenon strengthens the case that there is continuity between rule based and associative learning processes. Or conversely, that such results (e.g. Rumelhart and McClelland, 1986; Elman et al., 1996) show that neural networks are capable of rule-like behaviour. However because Pothos does not specify what rules are and assumes that they can operate on abstract representations, the review does not speak to issues of whether the seeming algebraic structure of language requires innate knowledge.

1.3.4. Peer commentary on Pothos. Pothos (2005) was accompanied by extensive invited peer commentary. This section very briefly reviews some of the criticisms made and the author’s responses. (Unless otherwise cited, all authors named in this section are commentators on Pothos. )

A number of commentators point out that ‘rules’ and especially ‘similarity’ are very flexible concepts that can be widely applied with a consequent loss of explanatory power. Pothos counters that this was a motivation for his review and attempt to pin down Rules and Similarity to particular set of cognitive operations. There is a further difficulty of determining a definition by which one could accept if a task is rule, similarity or intermediate. The mapping from this model to experimental data is also troublesome because of this lack of clarity, but Pothos views his approach as a way out of this more general problem of lack of definitions. Pothos does not disagree with Hahn that his proposal “turns what was evidence of the conceptual contrast between rules and similarity into their definition” (p. 25, Hahn).
Both Bailey and Marcus suggest that similarity involves comparison of two representations according to some appropriate metric, whilst a rule decision can be a computational operation on a single target representation (e.g. Does object process property X? Yes or No?). Similarly Smith suggests that similarity is a holistic process whilst rule are computed feature by feature. All these positions imply that is a rules/similarity dichotomy and that it is not the number of dimensions but how they are operated on that is important. Pothos gets round this in two ways. Firstly, he allows rules to operated on abstracted relational properties of properties and categories thus the specifically algebraic properties of rules that Marcus requires are to some extent assumed as part of a prior process. Secondly, rules may just be comparison to memory based exemplars that have built from many previous cases and this alone can explain the restriction to representations that are made up of only the relevant features.

Heit and Hayes and Ashby and Casale observe that individuals will spontaneously categorize by a single dimension whilst under time pressure individuals reason by analogy in preference to logic. Thus, in some cases rules are preferred but in others similarity takes precedence. Pothos suggests this depends on the coherence of the categories under consideration and that analysing category coherence under an organising principle of simplicity (e.g. Pothos and Chater, 2002) predicts when rules or similarity will be the default.

Cleeremans and Destrebecqz, Hampton and Smith all believe that the defining characteristics of rules is that they must be explicit. This was also the original position of Ashby that rules must be verbalizable but is given less emphasis in his later accounts. Pothos agrees that rules operations will often require explicitness and selective attention but he argues that this is a consequence of those cases where rules operate on a few properties of the target representation. Furthermore, his definition gets away from the problem of rules only being said to be present in those cases where participants can say that they are present (i.e. give verbal reports of rule use.) However, Ashby and Casale point out that Pothos (2005) does not address any of the empirical data outlined in section 1.2.1 above. In his response, Pothos suggests that a number of these results may in fact be accounted for by a difference between implicit and explicit learning.
1.3.5. Comparing Ashby & Pothos. Ashby and colleagues assemble a very comprehensive range of empirical evidence that there are multiple, qualitatively different category learning systems. Whilst Pothos reviews theoretical approaches to categorization and claims that these can be reassessed in terms of a continuity between rule and similarity processes. However, there is not necessarily any conflict between these two positions. There may be biases for certain types of processing in particular brain areas causing the specialisms that Ashby posits, but the main difference between them may be in terms of the number of stimulus dimensions they operate over.

Although focusing on category learning, neither Pothos nor Ashby addresses the importance of prior learning. There is likely to be a great deal of use of analogy and mappings to previously learned categories. Brooks and Hannah (2005), in their commentary on Pothos make the point that reasoning by analogy and the use of existing knowledge are key to a large proportion of rule use, and that these "instantiated rules" do not fit into the simple continuum of rule and similarity. This echoes Murphy’s view of concepts as the 'mental glue' that combines existing knowledge with new experiences (Murphy, 2002). This is crucial to any comprehensive theory of categorization because only very rarely will categorization take place outside the context of prior knowledge. As can be seen from the task illustrated in Figure 1.2(a). Although it only has two dimensions, judgments depend on information integration across both, because they cannot be separated, and therefore individuals tend to use a similarity type mechanism to solve this problem. There is no criteria in the Pothos formulation that would allow one to predict that although in this case only a 'small' absolute number of dimensions are used, the mechanism appears to be one of similarity. This appears to be another place where his model is under-specified.

In Ashby’s model there is competition between the rule and the similarity mechanisms that could account for apparent mixed performance. Pothos’s model does not permit this but instead allows processes that are intermediate between rule and similarity. Although these two explanations appear to be different, it is not clear without further analysis if they are in fact different. Nor is it clear what experimental data could decide between them.
Both these research perspectives start from the position that certain tasks are performed differently and seek to explain why a given task appears to be more rule-like or more similarity-like. For example, Ashby claims rule-based tasks are those which can be solved via explicit reasoning but does not elaborate this criterion. Whereas, although Pothos suggests that rules are equivalent to those cases where one attends to just one or a few psychological 'dimensions', it is left open what qualifies as a representational dimension and it is suggested that these may be quite abstract. Furthermore, quick survey of experimental results suggests that it can be problematic to determine when adults are using rule- or similarity-based mechanisms to make classifications.

1.3.6. Experimental problems with the Rules and Similarity distinction. In laboratory investigations of categorization, a free sorting procedure is typically used to assess adults' category learning. In these paradigms participants are presented with an array of cards, each of which features a single exemplar from a set that varies on a number of dimensions. Participants are required to sort these into two groups. Adults typically divide the items on the basis of a single dimension (Medin et al., 1987). Although similarity sorts are also found, the likelihood of a similarity sort appears to be increased by a range of factors such as stimulus complexity (Smith, 1981), integral (non-separable) dimensions (Handel and Imai, 1972; Garner, 1974), sequential presentation (Regehr and Brooks, 1995) or in speeded judgments (Smith and Kemler Nelson, 1984; Ward et al., 1986).

However, problems for any account of a simple contrast between rules and similarity have been raised by recent results from the adult literature which make the interpretation of adult similarity sorting less clear. In a series of experiments with adult participants using a match to standards paradigm, Milton and Wills (2004) found that contrary to Regehr and Brooks (1995), sequential sorting can lead to dimensional sorts. Whilst their first experiment replicated the Regehr and Brooks (1995) finding that participants sort by family resemblance when the stimuli were highly discriminable simple stimuli with low spatial integration, they found surprisingly that greater spatial integration lead to increased uni-dimensional sorting whilst changes in perceptual difficulty appeared to have no consistent effect. This
contrasts with early work which associated greater integration with increased family resemblance sorting (Handel and Imai, 1972; Garner, 1974). Milton and Wills interpreted their results as support for the idea that participants are using an analytic approach, perhaps based on a dimensional summation of stimulus properties. Milton et al. (2008) ran a further set of studies that supported this theory. They found that both increased time pressure or cognitive load from a dual task can decrease the likelihood of sorting by overall similarity. This suggests that overall similarity (at least in these particular types of laboratory tasks) is an analytic process.

To complicate the picture further a recent review of the comparative literature by Lea and Wills (2008) suggests that it is inappropriate to treat single dimensional responding as evidence for rule use and family resemblance sorting as evidence of a simpler associative processes. They contrast the experiments of Milton and Wills (2004) and Milton et al. (2008) showing that human similarity judgments can sometimes be based on analytic rules with research showing that pigeons can respond to single dimensions but do not (always) learn similarity based concepts. They observe that pigeons are highly capable of responding to single highly salient predictive features (Lea et al., 1993) but that in laboratory situations it can be very difficult to get them to learn artificially created polymorphous categories (Lea et al., 2006). Whereas classical results show that they easily learn complex natural categories such as the presence or absence of a human in a picture (Herrnstein and Loveland, 1964) which is presumably based on some similarity mechanism. While highlighting the complexity of the rules and similarity problem in human (adult) and pigeon studies Lea and Wills are not able to offer a full resolution of the issues. They do however make proposals that stimulus salience and depth of processing should be considered and that a highly empirical approach is needed where careful consideration is given to exact nature of the tasks and stimuli. An alternative approach might be to start from a set of defining characteristics of rules and thereby establish a set of criteria that can determine whether rules are being applied in a given context.
1.4. In cognition, what are the rules?

1.4.1. Aims. The following sections aim to survey instances of 'rule use' in various cognitive literatures, looking to see what, if any, definitions of rules have been used or assumed in each field. We do this to an attempt to provide a set of definitions that can be applied to rule use in the context of category learning and concept formation tasks. We will look at examples from the psychology of language, reasoning, executive function and animal cognition, as well as from the category learning literature itself. We will also briefly consider some issues raised by more general theoretical debates concerning classical cognitivist versus connectionist approaches to cognition.

We want to provide a definition of what rules are in order that we may clearly judge cases where they are said to be used. This is especially important when looking at the question from a developmental perspective. We wish to track the emergence of full adult-like performance using rules from its earlier possibly imperfect developmental precursors. If we can find a list of criteria that characterise rule use, we can potentially analyse infant performance in terms of the failure against some or all of these criteria. This will clarify our analysis of infant performance failures on some tasks. It will also provide a framework that might suggest infant experiments which rather than testing for full adult rule use independently assesses the competencies against each criteria.

Reviewing definitions of rule use will also help us to think critically about any instances of adult rule use that require special attention or explanation or deciding which tasks can be treated equivalently. We want a principled theoretical reason for expecting different performance on different types of task. In particular, we would like to be able to judge when a category learning task could be expected to require rule-use and when it might be more appropriate to classify it as a similarity based task.

1.4.2. Preliminary Considerations.

1.4.2.1. Three types of rule use. If we were to assume for the present that we knew what rules are, there are three ways in which human thought could be said to use rules;
1.4. IN COGNITION, WHAT ARE THE RULES?

(1) Rules as laws:- rules as either physical laws or of logical necessity.

(2) Rules as descriptive of behavior & cognition:- rules apply in a normative sense. They are statistical descriptions of underlying associative processes.

(3) Rules as actual mechanism of thought:- cognition operates by the manipulation of abstract symbols.

In the first case, the rules might be examples of physical or logical laws that thought must perforce be consistent with. An example of a rule of the first kind might be the application of the law of excluded middle; the observation of Aristotle that "there cannot be an intermediate between contradictories, but of one subject we must either affirm or deny any one predicate" (Metaphysics, Book IV, CH 7, p. 531). As such rules of this type are of least interest to us, because individuals are merely conforming to them rather than implementing them or demonstrating knowledge of them. Nonetheless, it is important to consider what rules might be of this type so that we do not over-interpret conformity as competence. (See section 1.6.1 for another example.) Furthermore, note that mere logical necessity is not sufficient for expecting rule compliance; all true statements in logic are tautological but very few are 'obvious' or immediately apprehensible. Clearly then there must limits on the complexity of the logically undeniable rules that people conform to. We might describe these rules as 'axiomatic' of thought.

In the second case, the rules might be descriptive of ideal or optimal human performance on a task and behaviour may approximate this whilst the actual mechanism is not itself rule based. An example of this kind might be found in psychological approximations to the law of large numbers; the statistical law that the mean of a sample approaches the mean of the population as the sample size increases. In their study of children’s understanding of chance, Piaget and Inhelder (1975) found even children as young as 10 years old were able to describe and utilise a heuristic that approximates this law. Whilst Nisbett et al. (1983) found this principle to be widely applied by adults. However, it is unlikely that individuals are explicitly using the rule directly. Similar considerations may apply to many other heuristics in reasoning. See Gigerenzer and Todd (1999).

Thirdly, the rules may be the actual mechanism that humans make use of in a given situation. For this to be the case, there should be a direct mapping between rule
and it’s mental representation. Examples of this kind of rule might be the ability to use the *modus ponens* form of logical deduction;

\[
\begin{align*}
\text{if } p \text{ (is true) then } q \text{ (is true),} \\
p \text{ (is true)} \\
\text{therefore } q \text{ (is true).}
\end{align*}
\]

Or the ability to correctly apply the concepts 'same' and 'different' correctly in a wide range of contexts. Here we will find that the rule is applied by the brain has exactly the same logical structure as we are formally able to describe it. (Although for the moment we will not address questions about the neural implementations that make this possible.) One consequence of this will be that formal analysis of the rule may lead us to certain expectations about applicability or implications and we can test these directly using appropriately constructed materials. For example, if the mind implements *modus ponens* in a formal and abstract sense then we would expect performance to equally good on questions of the form "Is q true given \(\{p \rightarrow q, p\}\)?" for any values of p and q.

It is important to keep these three alternatives in mind because they will have important implications when considering any example of rule use we choose to investigate experimentally. If the rule is axiomatic or the consequence of some physical constraint on the system then we might expect it to always be present. Furthermore, we should not give undue credit to the cognitive architecture for competence with this rule. In particular, we should not take a case of an infant’s conforming to a rule of this type as evidence for precocious cognitive ability to use rules. In the case where the rule is normative and descriptive of behaviour, we can expect that actual performance may match a rule to a greater or lesser extent.

Of course, for a given individual’s performance on a given task, we cannot say which type of rule use their behaviour follows. Nonetheless, it will always be important to keep the foregoing discussion in mind. We should guard against the over-attribution of abstract cognitive abilities to situations where low level or statistical mechanisms could equally well account for the performance. Hopefully, however, the following review will give us criteria for deciding between the alternatives. This is especially
true in the case of infancy where we may be looking at extremely simplified examples. The ability to measure performance against several independent criteria relatively independently will be useful.

Furthermore, in the third case of actual rule use, the rules may be either implicit or explicit. A distinction that is partially captured by the contrast between conscious and unconscious rules. In some cases individuals may be able to verbally report the rules they are using whilst in others the individuals are not consciously aware of the rules they are actually following, as for example in the case of most rules of grammar and syntax. Clark and Karmiloff-Smith (1993) point out that this distinction between mere rule-based behaviour and rule-like knowledge has important philosophical implications about the level at which representations can occur. A network that exhibits rule-like behaviour has "knowledge in the network" but not necessarily "knowledge for the network". It does not know that it knows. This is in contrast to the case of a conscious agent where some sort of meta-representations are apparent. Unfortunately, it is very difficult to distinguish experimentally between implicit or unconscious rule-like behaviours that are the consequence of rules implemented in a network and explicit or conscious knowledge of rules that operates on meta-representational level. For example, asking subjects to verbally report the rules they are using is an unreliable means of assessing their actual performance.

1.4.2.2. Knowing when to use which rule. Another subtlety to consider here is the interaction between rule use and rule identification. Consider trying to apply the size relation "bigger than" to two things A & B. We know from experience that this rule gives answers questions like "Is a house bigger than a car?" and "Is five bigger than six?" but not "Is five bigger than a house?" or "Is red bigger than green?" To apply the rule we first need to attend only to some particular size dimension of A and B, ignoring other features of these items and what is more it must be the same measure of size in both cases. Only then may the rule be applied. In some cases such as this one the rule can be applied at a perceptual level. In others the rule may operate on a more abstract level. For example, the question "Is number $x$ divisible by three?" can only be answered following the application of a sequence of abstract rules within arithmetic. Smith (1989a) speculates that it is the development of awareness of these "kinds of similarity" that "may be one of the major intellectual achievements of early childhood" (p.146, Smith, 1989). From
In cognition, what are the rules?

1.4

Figure 1.3: Schematic representation of the two phase model of rule based reasoning showing how it contrasts with exemplar based reasoning. In exemplar based reasoning, item pairs are directly associated with some result e.g. \{x_1, y_1\} -> 0, \{x_2, y_2\} -> 1, etc. By contrast, in rule based reasoning a process of abstraction maps items onto variables in an symbolic representation e.g. \{x_1, x_2, x_3, \ldots\} -> X, etc. Then the rule is applied at this abstracted level.

These examples we see that rules may operate with a greater or lesser degree of abstraction and that the process of abstraction is very dependent on the rule under consideration.

Smith et al. (1992), in a discussion of rule use in reasoning, suggests that rule use happens in two stages. First the individual must identify that a rule is applicable in a given situation and secondly they must apply it. Smith et al. (1992) refers to these two stages as abstraction and application. Others, who take issue with the idea of that human cognition uses rules, would say that a case-based reasoning system could achieve the same results without abstraction (e.g. Lewis, 1988). Whilst Barsalou (1990) argues that distinguishing between abstract rule-use and exemplar based account depends on how abstractly the problem is represented. Figure 1.3 illustrates this proposed two phase process and contrasts it with direct case-based reasoning; if there is little or no abstraction, rule like performance can be achieved with instance based learning. Alternatively, each variable element of the rule must be abstracted before a general abstract version of the rule is applied.
There are two related problems that our definition of rule-use should address. Firstly, at what level of abstraction the problem is being solved? Is there is some symbolic re-representation of the problem or can equivalent performance be attained with an exemplar-based approach. Secondly, is the process of abstraction itself rule based? In category learning, where we are interested in questions about category membership, it is arguable that there is no rule application stage and we are primarily interested in the process of abstraction that forms the category. The literature on category learning suggest that category membership can be determined by either rule or similarity-like mechanisms (see Ashby and Spiering, 2004; Pothos, 2005). Thus it is clear that in some cases the task may potentially require both rule and similarity processes. For example, the variables being abstracted by a similarity based categorization and then rule like operations performed at a conceptual level.

A further related issue is the danger that by separating the abstraction from the application stage we may map several different rules onto the same structure. For example, Cheng and Holyoak (1985) showed that the Wason card selection task (Wason, 1960) is much more easily solved in the 'permissions' case ("A person must be over 18 to drink beer") than the standard logical case ("If a card has a vowel on one side, it must have an even number on the other"). These versions of the task are formally identical but the difference in performance suggests there must be separate mechanisms for each type. In the following sections, we shall need to consider that apparently similar tasks may be solved by different mechanisms and also that several possible mechanisms could give rise to the same performance. This again emphasizes the importance of having a framework that defines rules and their application.

This brief consideration of philosophical and conceptual preliminaries to a discussion of rule use has been very far from exhaustive. The nature of rules in cognition and the structure of thought is an issue that has long been debated in the history of thought, forming a key point of contention in the debate between Rationalists and Empiricists. A debate that is kept alive by forceful advocacy of Jerry Fodor and colleagues (for example Fodor, 1980; Fodor and Pylyshyn, 1988; Fodor, 2000). Given this weight of history and continued controversy we begin with a survey of claims about rules use in reasoning.
The field of reasoning and rational cognition has long been concerned with how human thought processes can exhibit rule-like behaviour and if reason is governed by rules. This idea has a long history stretching back to ancient Greece and has been of interest to a wide range of philosophers and cognitive scientists since then. This section will briefly survey some elements of the history with the aim of understanding the assumptions that prevailed in earlier conceptions of human thought processes and indicating how and for what reason ideas have changed. After a short historical survey of cognitive science, we will present an overview of a couple of modern perspectives from Smith et al. (1992) and from Anderson’s ACT-R model (Anderson et al., 2004). The final part of this section will look at the work of Fodor and some of his critics.

1.5.1. Rules and Reason in the History of Philosophy. As far back as ancient Greece philosophers attempted to formalize reasoning and determine the rules underlying thought and logic. Aristotle (384 - 322 BCE) produced six volumes on logic collectively known as the *Organon*. These introduced many ideas for the first time, the first development of syllogistic reasoning and term logic (in *Prior Analytics* and *On Interpretation*), the introduction of hierarchies and an ontology of category theory (in *Categories*), the first attempt to develop and apply scientific methodology (in *Posterior Analytics*) and the first investigations of theories of formal and informal reasoning (in *Topics* and *On Sophistical Refutations*). Aristotle’s foundations of logic and his viewpoint that man was a “rational animal” would provide the dominant conception of reason in the Western European tradition for several millennia. The works of Euclid (c.325 - 265 BCE) were equally influential in the formation and maintenance of the idea of pre-eminence and comprehensiveness of calculation as pathway to knowledge. Euclid’s axiomatic approach to geometry provided a textbook case of how a complex and complete system of thought could be built up by the logical combination of simple self-evident first principles.

This is not to say that philosophers did not recognise the very large gap between, on one hand, statements about idealised abstract structures in the restricted field of plane geometry and on the other, the whole range of complex human thought processes about the real world. Nonetheless, the prevailing belief was that rational
principles were ultimately the cause and the mechanism of human thought. For example, in his book Leviathan published in 1651, Hobbes conceives of man as a calculator and reason as the process of calculation (reckoning):

\[\text{[w]e may define ... what that is which is meant by this word reason when we reckon it amongst the faculties of the mind. For reason, in this sense, is nothing but reckoning (that is, adding and subtracting) of the consequences of general names agreed upon for the marking and signifying of our thoughts; I say marking them, when we reckon by ourselves; and signifying, when we demonstrate or approve our reckonings to other men. (\S 5.2, Leviathan, Hobbes, 1651)}\]

A change in perspective and a challenge to this entrenched position came with the Empiricists. John Locke (1632-1704) promoted the view that ideas are built out of the association of sensations and reflections. Furthermore, all simple ideas (colour, form, number, etc.) proceed from perceptions. Locke did retain some rationalist assumptions, believing that the process of reflection was generative and therefore thought and its objects contained more than a mere association of primary sensations.

Whilst, David Hume (1711-1776), in his work A Treatise of Human Nature (1739), showed that there are severe limitations upon the power of inductive reason, upon the principle of causality and therefore on the conclusions one can draw about the world. In the opening sentences of this work he sets out a more radical skeptical position than that of Locke:

\[\text{All the perceptions of the human Mind resolve themselves into two distinct kinds, which I shall call Impressions and Ideas. The difference betwixt these consists in the degree and force and liveness with which they strike upon the mind, and make their way into our thought or consciousness. (Book I, Part i, Hume, 1739)}\]

Hume’s position is therefore highly skeptical and a profound challenge to all aspects of a Rationalist world view. In particular, by showing the limitations of deductive and inductive reasoning he radically limits what can be said to be true about the
For Immanuel Kant (1724-1804), it was reading Hume’s works that ‘woke him from his dogmatic slumbers’. He could not accept Hume’s conclusions and worked on a riposte. The result was his *Critique of Pure Reason* (1781), a work equally as long as Hume’s *Treatise* but which came to very different conclusions. The argument was dense, technical and hard to summarise but ultimately Kant’s position is one of transcendental idealism. He believes that certain *a priori* features of the world and the mind make the former comprehensible to the latter. He relates what we can know to how we can know it. We can never know the world as it is in itself but we can go beyond the mere perception of *phenomena* because of the conceptual framework provided by logical synthetic reason; that is reason about perception rather than pure reason. In Kant’s words, "thoughts without content [are] empty, and intuitions without concepts [are] blind". This has been taken as psychologically nativist stance that requires the mind to have logic, reason and rules built in before it can understand anything about percepts.

Part of Kant’s argument was built on the self-evident ‘truth’ of Euclidean geometry and Newtonian physics. Thus, the discovery, at the end of the nineteenth century and the turn of the twentieth, of non-Euclidean geometries and non-Newtonian physics undermined his position somewhat but the essence of his ‘intuitive’ framework remains. We cannot conceive of objects except in time and space, even if (as it turns out) we do not know exactly what time and space are.

There is one further philosopher whose thoughts about rules in thought ought to be considered, namely Ludwig Wittgenstein. While Kant changed the focus from trying to know how the world actually is to understanding how it seems to be, Wittgenstein went one step further and considered only questions about what we meaningfully *say* about the world. His first work, the *Tractatus Logico-Philosophicus* (1922) was an attempt to show thought consists of statements in language that correspond to facts about the world and that philosophy consists of nothing more than the analysis of the logical form of these thoughts and sentences.
However, his later work was a direct challenge to his earlier position. For the bulk of his second major work *Philosophical Investigations* (1953), he argues against the idea that there can be an internal private language of thought. Which has the consequence that all language is in some sense public. The meaning of words are a set of socially agreed and transacted usages. Wittgenstein illustrates this showing how difficult it is to give a definition of the word "game". Any such definition must cover the word’s usage as it understood when referring to a game of chess, of football, of solitaire, or of catch. No such definition is forthcoming, yet we understand how to use the word without knowing any perfect definition of it. This is not possible for any private language because there are no external referents and no way of confirming the meaning attached to an word. It is the interpretation of Saul Kripke (1982) that Wittgenstein’s rules paradox is the key to this skepticism about private language. He considers this to be a direct consequence of Wittgenstein’s proof that no rule can be known to apply to an action:

This was our paradox: no course of action could be determined by a rule, because every course of action can be made out to accord with the rule. The answer was: if everything can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here. (§201, Wittgenstein, 1953)

Although we may continue the sequence \{2, 4, 6, 8, \ldots\} by applying the rule \(x_{i+1} = x_i + 2\), the sequence so far also accords with many other possible rules. Thus, the concept of a rule is under-determined. Dror and Dascal (1997) argue that Wittgenstein’s paradox might provide a principled philosophical reason for rejecting the classical symbol processing model of cognitive science in favour of a rule-free connectionist approach. However, Wittgenstein is a notoriously difficult philosopher to interpret and there is considerable disagreement amongst philosophers about the meaning and implications of his rules paradox. Fortunately, we are not directly concerned with the philosophical foundations of cognitive science. We, therefore, proceed to investigate rules as envisioned in classical cognitive science.
1.5.2. The Cognitive Revolution. In the first half of the twentieth century, behaviourism dominated experimental psychology and mental 'rule use' was dismissed as an irrelevance, unprovable and inaccessible. Behaviour could be explained in terms of stimulus-response associations and operant conditioning. However, in the post war period there was resurgence of interest in rational approach to cognition. In part this was due to Noam Chomsky’s devastating critique of behaviourism (Chomsky, 1959) and his equally revolutionary program in linguistics (Chomsky, 1957). Equally important was the reaction to the emerging technology of electronic computation and the birth of the classical approach to Artificial Intelligence (AI) which sought to emulate human-like intelligent behaviour using programmable computers. Alan Turing’s formal theory of universal computation (Turing, 1936) and its first implementations in electro-mechanical and electronic computers led to the beginning of computer science. Whilst McCulloch and Pitts (1943) speculated that neural computations in the brain were equivalent to the logical operations of AND, OR, etc gates of computer science.

New computer technology brought novelty and promise of its own and it also brought new metaphors for describing the mechanics of thought-like processes, memory systems and self-regulation in machines. Inevitably these became the new models for describing human thought. In some cases researchers went further and built computer simulations of human reasoning according to the principles of computer science. Of necessity these operated according to logical and procedural rules and carried an assumption that the brain must implement some analogous mechanisms. In fact, it was the specific aim of this type of research to elucidate mental function through a precise (and psychologically plausible) reimplementation of rational cognition. This was in contrast to much of artificial intelligence research which was interested in pushing the boundaries of machine intelligence by whatever means technology and algorithms could provide. Naturally, this often involved looking to the human model for inspiration, as in much of early computer vision research (see Chapter 1 in Boden, 1988). But it is only in the case of attempts to precisely simulate certain rational behaviours where the real measure of success is how the computer model aids our understanding of the actual processes of human thought and in particular the rules it obeys or implements. The next two sections describe two such models that are typical of this approach; the highly influential
1.5. RULES IN REASONING

General Problem Solver of Newell and Simon (1963) and John Anderson’s ACT-R model.

1.5.3. General Problem Solver. Newell and Simon’s (1961) General Problem Solver was a standard procedural computer program that was designed to produce serial, step by step solutions to logical problems. Although they restricted their analysis to problems with a very restricted and simple logical structure, Newell and Simon presented their GPS as a simulation of high level cognitive abilities. It was their ambition that the framework they were introducing was, in principle, very general. From the outset this claim was controversial. Some critics claimed that the whole approach was fundamentally misdirected because thought is parallel and associative rather than procedural. Whilst others accepted the principle but questioned the generality of the particular framework that GPS represented. Nevertheless, Newell and Simon’s work was extremely influential both for the theory and concepts it introduced and for their methodological approach which tied the models and programs very closely to real experimental data and simulations, requiring the one to inform the other directly.

How GPS and production systems work and what they assume and imply about rational cognition is best illustrated with an example. Indeed, the first step of Newell and Simon’s approach was a detailed task analysis, examining the actual performance of experimental subjects on a specific class of logical problems. One of the tasks that Newell and Simon analysed in their book Human Problem Solving (1972) was the solving of crypt-arithmetic puzzles. Consider the simple sum shown in 1.1.

\[
\begin{array}{c}
B & E \\
+ & B & E \\
\hline
S & E & E
\end{array}
\]

(1.1)

Each letter represents a single digit and the object is to find the unique solution that makes sense mathematically. In this case, the process is very straightforward. \( S \) must equal 1, since no larger carry is possible from the sum of two digits. This also implies that \( B \geq 5 \) in order for there to be a carry. We could try some values for
$B$, but without knowing $E$ this will be indecisive. Therefore we turn our attention to the right most column. We might see that this implies that $E$ must at the very least be even. Trying \{2, 4, 6, 8\} each leads to a contradiction, and this gives the full solution \{S ← 1, B ← 5, E ← 0\}.

Newell and Simon start with this type of descriptive account of the reasoning process and try to formalize all aspects of it. The first thing they define is the problem space itself. In this case, the set of letter-number substitutions and their equivalent arithmetic with the goal of finding a consistent solution. As in the human reasoner, GPS would work towards this goal via a series of sub-goals (e.g. finding the value of a single letter). It worked towards these using a set of operators, rules it knew how to apply. The majority of the rules of the system are of the "if $p$ then $q$, else $r$" variety but they also can be more complex. Notice that even simple cases require the ability to check the truth of statement $p$, which could in itself be a complex operation. For example, in this case the problem solver must have a concept of 'even'. Although equally some truths can just be known, for example the fact $5 + 5 = 10$ could be retrieved from memory. The chaining and branching of if...then..else... constructions is characteristic of procedural programming, so the GPS has the full power of a Universal Turing machine. Even a fairly small set of operators can lead to rapidly combinatoric expansion of the problem space. Therefore, Newell and Simon also included heuristics that would hopefully guide the problem solver rapidly toward the solution, whatever the problem space. For example, backtracking efficiently from contradictions and testing cases that reduce the remaining possibilities by the largest amount. Backtracking and exploration are possible thanks to a goal stack that keeps track of order in which it applies steps.

By taking this approach, Newell and Simon were able to model problem-solving tasks such as the one above and even some basic motor skills (Card et al., 1983). The models fitted participant data to impressive degree of precision, including in some cases, measures of the hesitations, introspection and eye and hand movements. GPS never fulfilled its ambition of becoming a general framework, although did it inspire the majority of the later models and methodologies of rational cognition including, as we shall see in the next section, one of the most comprehensive and successful models of in the field. We will save our assessment of what these types of approaches
have to save about the nature of rules and their usages until the next section. Before passing on it is worth noting some high level assumptions about GPS. Operation was strictly serial and deterministic, all steps took place one after the other without error in a central processor. Similarly, it is a strictly symbolic system, all tokens can be passed between all procedures or held as chunks in memory, but tokens lack family or category structure. Finally, it is non-developmental nor even a learning model, competence is fixed by the rules it is provided with. With the benefit of hindsight, these are all grave limitations of the model, but GPS’s innovation and performance impressive in terms of what came before and the models they inspired. We therefore turn Anderson’s highly ambitious and extensively applied model ACT-R, which might be thought of as a modern descendant of the stone age GPS.

1.5.4. Anderson’s ACT-R Model. This section briefly reviews Anderson’s Adaptive Control of Thought - Rational (ACT-R) model. This model has gone through numerous incarnations and refinements (Anderson, 1983, 1993; Anderson and Lebiere, 1998; Anderson et al., 2004). Here we will concern ourselves with the version presented in Anderson et al. (2004) although the exact details of implementation and architecture are of less interest than the principles motivating the model.

ACT-R (version 5.0) is a model of human knowledge and reasoning, both in use and acquisition. It owes a great deal to the pioneering architecture of GPS but with many crucial modifications. As can be seen from figure 1.4, ACT-R is a highly modular architecture. It comprises four distinctive systems represented by the visual, motor, memory and intention/goal modules, which are co-ordinated and integrated by a central production unit. From our perspective the visual and the manual modules which determine how the model perceives and interacts with the world are the least interesting. Similarly, the memory modules just store information, the nature or format of which will be determined by what the rest of the system provides and utilises from memory. Most relevant to our concerns our the intentional module and the central production module, the action and architecture of both of these determining how rules are embodied in the system.

At its main level of operation it is a symbolic system but it has sub-symbolic aspects. Notably, it allows an associationist architecture for its memory system. It
has two types of long term memory procedural and declarative. Search and retrieval of these can operate in parallel but the use of the information is restricted to serial processing by a buffer on retrieval. It also possesses a goal stack which is a form of working memory. It aims to be consistent with neuroscience theories of brain area function. For example, the goal module is associated with the dorsolateral prefrontal cortex (DLPFC) and memory is associated with the hippocampus. Similarly, it tries to abide by neurophysiological constraints on architecture and performance. Neurobiological evidence suggests that the basal ganglia is associated with rule use (see Ashby and Spiering, 2004). In this way ACT-R aims to be informed by neurobiology but is short of being a complete or final model. As is recognized by the authors; "the proposal that all neural processing passes through the basal ganglia systems is simply false" (p.1057, Anderson et al., 2004).

Two ways that ACT-R improves on GPS are by being probabilistic in its choice between operators and by being able to learn. At a given decision point, it weights
each possible production by a utility, which combines a probability of success and a
cost in terms of how many more steps it should take. It always chooses the action
with the highest utility but with experience it adjusts and improves its probability
and cost estimates for that choice. Furthermore, by starting with appropriate
unbiased estimators this process is approximately Bayesian. However, notice that
it is probabilistic only in terms of rule selection, not in terms of the rules themselves.
These have the same if...then...else... structure as GPS and like in GPS they are all
pre-set before the model is run against a particular problem space. This highlights
that ACT-R can model learning (choosing the right rule) but not development
(creating new rules), all performance is in terms of a mature system, where rules
are acquired in an all or none manner (albeit they are then applied probabilistically).
There is no mechanism for rules to be emergent or for any type of 'representational
re description (Karmiloff-Smith, 1992).

Nevertheless, the ACT-R architecture has been applied to very large range of do-
 mains and problems and has a large following in the research community, as can
been seen from the detailed examples modeling real performance of anti-aircraft
warfare co-ordinators and naive problem solving in a toy-algebra in Anderson et al.
can give remarkable fits to real data without relying on multiple free parameters,
that is genuine prediction not model-fitting post-diction. While it is a modular
system, Anderson and other researchers using ACT-R place a strong emphasis on
its ability to integrate cognitive, memory and sensorimotor systems. The model
has the ability to simulate all aspects of a behavioural task from visual attention,
memory retrieval and decision making through to sensorimotor output required to
give a response. This they argue makes its a much more powerful and predictive
tool for modelling actual performance data. This echoes the ambition of Newell
and Simon to be lead by pragmatic and practical appreciation of real data rather
than elaborate and distant theorizing.

However, a consequence of this integrative system level approach is that the focus
is on the higher level and as long as the performance is acceptable, the foundational
elements are taken for granted. The 'atomic components of thought’ are left un-
examined. The individual productions, the goal states, the chunks of memory are
used without internal examination of their epistemic state. The symbolic nature
of the system and its rational operation are assumed, whilst the parameters and architecture of the sub-symbolic aspects are primarily motivated by data fitting (for example, Anderson and Reder, 1999). The aim of this review is not to assess the ACT-R approach as whole, but rather to investigate its assumptions about rules in cognition. Here we have seen that it takes a classical approach, with a very symbolic and procedural approach to rules and their use, without examining their internal structure. We therefore turn to a paper that aimed to review rule use in the human reasoning literature and attempted to determine criteria by which rule use could be judged.

1.5.5. Smith, Langston and Nisbett’s Eight Criteria. Smith et al. (1992) argue that there is considerable debate and controversy over if and how rules are used in reasoning. Therefore, they present a review of the use of rules in reasoning that proposes eight criteria by which to judge if a person was in fact using a rule (Box 1). If someone really is using the rule in question then its application should meet all eight criteria. Furthermore, by explicitly providing a set of criteria, they hope to go beyond a simple dichotomy between rule-based and case-based reasoning. Depending on how many criteria a given behaviour matches the more or less rule-like it may be expected to be. In this section we will examine Smith et al.’s following their own grouping of linguistic, performance and training criteria.

The first three criteria are taken from psycholinguistics where there is much interest in whether a system has rules. But, Smith et al. argue, they are applicable more widely than that. The first two criteria approach the issue of abstractness. As we remarked in section 1.4.2.2 the main theoretical difference between rule-based and case-based is to do with the level of abstraction. These two criteria look for the behavioural consequences of a greater level of abstraction, arguing that it would imply equal accuracy with less familiar (criterion 1) or more abstract (criterion 2) material. Certainly this would not be the case for instance based reasoning, where greater similarity of a new item to already encountered exemplars would equate to better performance. But notice that specific claim is about accuracy, not overall performance as might be measured by taking account of both speed and accuracy. Recalling also from 1.4.2.2 that rule use has two phases abstraction and application, we see that if these were claims about performance, they would
Criteria Stemming from Linguistics

(1) Performance on rule-governed items is as accurate with unfamiliar as with familiar material.
(2) Performance on rule-governed items is as accurate with abstract as with concrete material.
(3) Early in acquisition, a rule may be applied to an exception (the rule is overextended).

Performance Criteria

(4) Performance on a rule-governed item or problem deteriorates as a function of the number of rules that are required for solving the problem.
(5) Performance on a rule-governed item is facilitated when preceded by another item based on the same rule (application of a rule primes its subsequent use).
(6) A rule, or components of it, may be mentioned in a verbal protocol.

Training Criteria

(7) Performance on a specific rule-governed problem is improved by training on abstract versions of the rule.
(8) Performance on problems in a particular domain is improved as much by training on problems outside the domain as on problems within it, as long as the problems are based on the same rule.

p. 7, Smith et al. (1992)

<table>
<thead>
<tr>
<th>Box 1: Eight Criteria for Rule Use (Smith et al., 1992)</th>
</tr>
</thead>
</table>

be strong criteria, implying that the whole process was comparable with familiar, unfamiliar and abstract material. In fact this requires the assumption that the abstraction phase itself is unaffected by familiarity. Smith et al. acknowledge this assumption but do little to justify or investigate it. A counter-argument might suggest that criterion 1 is confounded by the lack of a way to compare the costs of matching an unfamiliar item to an familiar instance (case-based) to the cost of performing an abstraction with unfamiliar material (rule-based). Even when dealing directly with abstract material (criterion 2) one is faced with an equivalent of this dilemma because there is a cost to knowing that a rule is applicable to a given type of abstract material (e.g. the relation bigger than only applies to size dimensions.) We return to the issue of abstractness in section 1.8.2 below.

The final criterion taken from linguistics is the over-extension of rule-use in early acquisition. Smith et al. have the specific case of the past tense in English in mind. We cover this in more detail in section 1.6.1 below where we see that this criterion need not necessarily support a rule based interpretation. Smith et al. do not provide any other strong examples of this criterion in action. Furthermore, we speculate that both mental set (Jersild, 1927) and memory schema (Bartlett, 1932) provide accounts of over-extension in instance based reasoning and therefore
It is not clear that either of Smith et al.’s performance criteria 4 or 5 are \textit{diagnostic} of rule use in a general sense because although both effects might be expected in cases of rule use, they could equally well be found with other cognitive processes. Priming is a familiar effect in memory and semantic tasks whilst complexity is synonymous with difficulty in numerous tasks. Nonetheless, the criteria could help assess rule use in certain circumstances. For example, if performance decreased linearly with number of rules. Although such effects could disappear with practice, if, as is possible in in Anderson’s ACT-R model, regularly encountered compounds could be ‘compiled’ into a single rule. Priming evidence would only be accepted when it occurred in the face of conflicting surface features or other potential primes. Smith et al. give an (unpublished) example from their own work where permission and obligation rules were primed by rules of the same type but not by each other. This result was confounded by potential priming by surface similarity (e.g. use of word ’can’ in both permission rules).

Criterion 6 is equivalent to the claim that the description a person gives of their reasoning in a given task \textit{may} contain an accurate report of the rules they did actually use. This has some similarity to the criterion given in Ashby et al. (1998) that rules are precisely those judgments that can be verbalized. Criterion 6 is very weak because clearly people can often use rules they are unable to describe (e.g. rules of grammar) whilst in other cases rules are reported when they are not actually used. In fact, it may be that the requirement of having to report something verbally leads experimental participants to frame relatively simple rule-like explanations for their behaviour. Nonetheless, protocol analysis is a valuable tool for assessing what rules might be used in a given situation, as long as there are independent criteria for judging if these rules are actually used. For example, in an artificial grammar learning task, Dulany et al. (1984) found that individuals’ self-reported rules could account for their accuracy on the task, although these sets of ‘micro-rules’ were often very limited in scope and validity.

Criteria 7 and 8 are again addressed to the abstract nature of rules. Criterion 7 claiming that rules can be learnt in the abstract and this will improve performance on specific instances. Whilst criterion 8 takes the same claim a step further and
posits that experience with a rule in a specific domain is actually improving the ability at the abstract level and hence (by criterion 7) will transfer to a new domain. Both of these criteria seem to be useful and diagnostic because they test consequences of the aspects that are crucial to any definition of rules, abstractness and generalizable. For both criteria, they cite a study by Fong et al. (1986) in which training with abstract statistical concepts (population, sample, variability, etc) or training on statistical problems in certain contexts (random number generators, population sampling or mean estimation) transferred to the other untrained domains and to problems about everyday life. This would seem to support the idea that participants are acquiring abstract rule based knowledge.

Taken individually none of Smith et al.'s 8 criteria is without problems and none is perfectly diagnostic of rule use. But each has at least some applicability and validity, such that taken together they can provide converging support for performance being rule-based, although other possible explanations may also fit the criteria. Notice, however, that they are quite strong criteria for deciding that a behaviour is not rule-based. If, for example, performance on particular reasoning task does not transfer to a logically equivalent domain or if adding complexity does not reduce performance then we could hypothesize that the underlying behaviour is nor rule based (while remaining aware of the dangers of drawing conclusions from negative evidence.) Thus, overall Smith’s review is useful towards our goal of characterizing the nature of rules and guiding empirical work that will assess development of rule use in infants and children. We now move away from empirical and pragmatic measures of rule use to a highly philosophical approach to the problem.

1.5.6. Rules and the 'Language of Thought'. In their influential article in Cognition, Fodor and Pylyshyn (1988) (hence forth F&P) make the case that all thought takes place at a symbolic level. Namely, that all thought is the result of symbolic representations that are manipulated according to syntactic and semantic rules which effectively form a 'language of thought'. They contrast this position with connectionist approach which they characterise by its supposed claims that such systems 'can exhibit intelligent behaviour without storing, retrieving or otherwise operating on structured symbolic expressions' (p.5, ibid.) The implication of this position is that all thought is ultimately rule-based. In this section, we
present a brief summary of F&P’s arguments, followed by some of the criticisms it provoked.

Crucial to F&P’s analysis is a concern over levels of explanation. They allow that connectionist architecture could implement a cognitive system with syntactic and semantic properties. Indeed, they consider the human brain to be just such an implementation. But they contend, at the computational level, it can only do so by implementing the equivalent of a Turing architecture (Turing, 1936). Thus their position rests on two claims. Firstly that it is possible to separate the algorithmic and implementational levels involved in cognition and secondly that only a Classical, rule-based system has the properties at the algorithmic level to do produce the productivity, systematicity and compositionality characteristic of human thought.

F&P address the separation of the algorithmic and implementational levels at the beginning and end of their article. Initially, they are concerned with showing that both connectionist and classical models of cognition must be Representationalist. That is, ‘there are states of the mind which function to encode states of the world’ (p.7, ibid.) In this case, then both approaches entail cognition to be concerned with the manipulation of internal representations (or symbols) that have semantic content. The second step of their argument aims to show that only when these representations also have syntactic properties can human cognitive performance be achieved. Finally, they revisit the separation of levels by dismissing the ‘Lures of Connectionism’ (F&P, Section 4), the idea that Connectionist but not Classical architectures provide a biologically plausible model of the very fast, massively parallel, noise and damage resistant processing characteristic of the brain. F&P dismiss this type of reasoning as too dependent on the limited metaphor of current capabilities of electronic computers, pointing out correctly that there is nothing intrinsic in a Classical rule system that prevents it from implementing probabilistic, noise resistant rules that can operate in parallelized system, respecting constraints like the 100-step rule. The continued absence of such classical models of cognition does not, by default, speak in favour of Connectionism.

More principled, theoretical arguments are required to decide between the two approaches and the bulk of F&P’s article is taken up with the goal. Arguing that only Classical systems can model cognition because only the mental representations
in a classical system can have *combinatoric syntax and semantics* and that mental operations are *sensitive to structure of the representations*, independent of content (i.e. only a classical system can have truly abstract representations.) They illustrate the need for syntax by pointing out that *John loves Mary* is a different thought to *Mary loves John* and so cannot be captured just by the simultaneous activation of representations for JOHN, LOVES and MARY. More generally, they claim that no set of atomic properties however they are combined is going to capture concepts of greater complexity unless there is syntax to give meaning to their combination. And conversely, given a syntactic structure, the meaning can be understood whatever particular variables fill the structure, understanding *John loves Mary* implies being able to understand *Mary loves John* and *Bill loves Jane*.

F&P go on to claim that only Classical systems are able to demonstrate productivity, systematicity, compositionality and inferential coherence. However, their arguments in this section are not convincing. They compare a hypothetical Classical system which by definition has these properties built in to a caricature of the contemporary (1988) state of the art in Connectionist models. They largely dismiss the possibility of a Connectionist system showing any of these properties on the basis that Connectionist architectures are mere statistical associative engines that cannot operate in an entirely abstract domain. In the theoretical limit this is true, but F&P do not present any evidence that Connectionist systems cannot approximate these abilities nor any corresponding evidence that human cognition is not itself constrained by experience. Even if one accepts the centrality of these properties to cognition, this seems to be more of an empirical question.

Many other authors answered the challenge to (modern) connectionism set forth by F&P. Chalmers (1990) suggested that F&P make the mistake of focusing on the atomic properties of localist networks and under-estimate the contrast with the distributed representations that give the distributed networks greater power and better correspondence to the messy category structures we encounter in the world. Related to this, there is the 'symbol grounding problem’ (Harnad, 1990). Even if we can successfully create a system that manipulates symbols, this does not require a correspondence between the symbols and what they represent in the world. In Harnad’s words one cannot "learn Chinese from a Chinese/Chinese dictionary alone’ (p. 335, Harnad, 1990).
Chater and Oaksford (1990) argue that F&P are wrong to assume that connectionist models do not seek to capture structured representations. Additionally, Chater and Oaksford identify an assumption of F&P about the autonomy of psychological explanations (i.e. 'thought') and implementational explanations (i.e. 'computation'), they show that this independence of levels of explanation cannot be maintained even in F&P’s own account of the 'lures' of connectionism.

One further issue that comes out of a consideration of the 'language of thought' concerns the definition of the symbols which the cognitive apparatus is said to operate on. Understanding what these symbols might be is not straight-forward. According to Alan Newell we can think of symbols as the mechanism that "provide distal access to knowledge-bearing structures that are located physically elsewhere within the system.” (Newell 1990, p.457) Here we may think of an analogy to variables and 'pointers' in modern programming language like C/C++. An ordinary variable is directly identified with an area of memory allocated for it’s storage whilst a pointer is a (smaller) variable that holds the memory address of some other variable and thus 'points to' or represents the it without requiring a full copy to be made. Operations can then be efficiently performed 'at a distance' using the pointer. This architecture is easy to implement in modern computer where memory is inert and uniform but more problematic in neural terms where memory is heterogeneous, hierarchical, semantically rich and perceptually embedded. Evidence from scanning studies suggests that there is substantial activation in supposedly perceptual regions when memory or imagination are invoked (Kosslyn, 1994), this counts against the idea that thought is symbolic in these instances.

Finally, Fodor and Pylyshyn’s 'Language of Thought' hypothesis does not offer much help towards our goal of characterizing rule-like behaviours and their development. If all thought is considered to be rule based at some level then there is little guidance as to the defining features that make some behaviours rule-like. Similarly, there is no developmental aspect to their theory. It is unlikely that they would claim that a language of thought is completely innate and does not, like verbal language, need to be learnt slowly and effortfully over the first few years of life. However, they make no speculations about this. More generally, their discussion remains through out on a very 'conceptual' level and makes no attempt to consider the implementation or empirical implications of their position, as such it is of little
1.6. Rules in Language

Language is a domain where several types of rules are said to operate. Language rules are divided into rules of morphology, syntax, phonetics and semantics. Morphology concerns the patterns and regularities in the way words and word stems are transformed with prefixes, suffixes, case and tense changes and other types of compounding and transformation. Syntax is the set of conventions that are followed in the construction of phrases and sentences. For example, what clause agreements are necessary or what orderings of Subject, Verb, Object are permissible or make sense in a given language. Phonetics addresses the way the words are transformed into sounds. Which phonemes the language uses and what combinations of sounds and syllables are permissible. For example, in French no word may begin with two consonants. Phonetics concerns how words are pronounced and how the language handles elisions and co-articulations, together with conventions about patterns of stress and intonation. Finally, semantics is concerned with how the language describes the world itself, with the meaning of words. It brings world knowledge into the purely symbolic domain of words and grammar. Semantics disambiguate homophones and the meaning of words in different contexts. It describes how idiomatic usage can give phrases a different meaning to their 'literal' meaning. Note also that, the rules of language are not explicit. Individuals make use of them without the need for conscious awareness of the rule itself.

In Words and Rules (2000), Steven Pinker lists three features of the rules of combinatorial grammar systems as typified by the various incarnations of the Chomskian program (Chomsky 1957; 1965; 1993). Rules are productive. By using rules, an essentially infinite number of new sentences can be constructed that have the structures and strictures of the rule system with new combinations of words. Rules are symbolic, which is equivalent to rules being abstract. They are defined and operate on a general level, defined at a remove from individual words or phonemes. For example, a rule will apply to the past tense of all regular verbs or to the word order...
of all sentences containing a subject, an object and a verb. Thirdly, rules are *combinatorial*. Several rules can applied sequentially or in conjunction with one and other. Including situations applying the same rule repeatedly, such as when one noun phrase is embedded in another noun phrase, leaving other rules unaffected.

For example,

NP

`Det N`  
|   |   |  
The dog  

and

NP

`Det NP`  
|   |   |  
The NP who VP  
|   |   |  
large black N V N  
|   |   |  
dog V N  
|   |   |  
likes N  
|   |   |  
sausages

are treated as grammatically equivalent noun phrases (NP) for the purposes of embedding in larger sentences. This property of *recursion* important aspect of this combinatorial feature of rules in language is that the rules are *recursive*. It has been suggested by Hauser et al. (2002) that recursion is one of the defining, special features of the uniquely human faculty of language. However, the rule based nature of language is by no means simple, uncontroversial or universally accepted.

### 1.6.1. The past tense in English.

The debate of the past tense in English is a particularly illustrative example here. In English the standard past tense form of a regular verb is formed by the addition of the `-ed` suffix. Thus *call* becomes *called*, *walk* becomes *walked*, *fit* becomes *fitted* and so on². This pattern applies to the vast majority of verbs in English. There are only about 150 to 180 exceptions. (The

²Although there are three different pronunciations depending on the final phoneme: `-d` as in called, `-t` as in walked and `-id` as in fitted.)
The initial wholly rule-based model of Chomsky and Halle aimed to go further than just accounting for the distribution of past tense forms. Their work *The Sound Pat-
tern of English attempted to explain the pronunciation and production of English words in all contexts using a small number of phonological rules. Furthermore, they claimed that this generative morphology is the set of actual rules that the mind uses. Thus, the sing-sang and sit-sat are a consequence of the 'Lowering Ablaut' rule, by which tongue height of vowels is altered. Whilst a vowel shortening rule, 'Shortening Ablaut' would account for flee-fled and shoot-shot. However, as Pinker (2000) points out, this latter rule does not work directly on modern English pronunciations but only makes sense in the historical context of the 'Great Vowel Shift' of the fifteenth century. Chomsky and Halle’s theory anticipates this by claiming the system operates by storing words in a highly abstracted underlying form which is then operated on by morphological rules to produce the spoken forms we are familiar with. Thus making it an extreme example of a rule system in terms of both abstraction and application.

Various empirical data provided problems for the Chomskian account. Foremost amongst these was the finding of Bybee and Slobin (1982) that irregular past tense of verbs cluster into family resemblance classes, groups of irregular verbs that have more or less surface similarity and follow the same transformation in the past tense. The membership of a particular family did not follow a clear rule but depended on the similarity of those verbs to others in the same cluster. For example, one regular group comprises verbs like rise, write, ride, stride, dive, thrive, shine, strive, etc. whilst a messier group includes blow, grow, throw, draw, fly, slay, etc. The Chomsky, Halle and Mohanan model makes no account of this similarity, treating each irregulars individually. Similarly, their model makes no predictions about word frequency or familiarity effects. Nor does it offer a developmental account of how children acquire knowledge of the underlying form, which in some cases they will never hear. For example, the theory claims that the underlying form of run-ran-run is said to be rín by analogy to sing in sing-sang-sung.

Connectionist models are good at handling similarity and frequency effects and by virtue to their learning regimes, they are also fundamentally developmental. The Rumelhart and McClelland (1986) model was proposed as an alternative parallel distributed processor (PDP) implementation of English past tense acquisition. Figure 1.5(a) provides a highly schematic representation of one version of this model. The model takes the verb stem as an input and attempts to produce the past-tense
1.6. RULES IN LANGUAGE

(a) Rumelhart-McClelland PDP model

(b) Pinker’s Words and Rules model

Figure 1.5: Two models of English past tense formation. See section 1.6.1 for details. Diagrams taken from McClelland and Patterson (2002b) and Pinker and Ullman (2002b) respectively.

form. Initially, the network outputs are random but it is trained by making small corrections to its internal weights and thresholds on each presentation of a training example. The training set consisted of 420 verbs of which a proportion were irregular. The single integrated network learned to produce all past tense forms correctly. During learning the model showed the U-shaped development found in young children where an initially correct irregular past tense form is temporarily supplanted by a regular + -ed form before finally correct form is e.g. give -> gave / *gived / gave.

Next the network was tested with 86 new verbs. These were a mixture of regular and irregular. The model gave the correct -ed past tense for most of the regular forms and made reasonable over-generalization errors for some irregulars such as caught or digged. Whilst it successfully predicted irregular past tenses for verbs in groups it was familiar with including cling -> clung, bid -> bid. However, the model had limitations, for several irregulars it did give any output and made some very unhumanlike errors, such as parsing mail -> membled. Pinker and Prince (1988) gave a comprehensive critique of the original Rumelhart and McClelland model. They pointed out the unidirectional nature of the model, the short-comings of the "Wickelfeature" representation of phonological features and the models inability to deal with homophones, compound words and other exceptions that were mostly
consequence of words being treated as semantic tokens and symbols by humans but as mere associations by the model. Rumelhart and McClelland were aware that their model was not a definite solution to what is a very complex problem, but presented it as a demonstration of the power of purely associative mechanisms. McClelland and Patterson (2002a) indicate that continuing work aims to extend the original idea based on a more recent connectionist model (Joanisse and Seidenberg, 1999) includes semantic information.

An alternative is the Words and Rules theory of Pinker and colleagues (Pinker and Prince, 1988; Pinker, 1991, 2000, 2001, 2002). Figure 1.5(b) illustrates the Words and Rules model. It consists of two competing systems, a symbolic grammatical module and an associative lexicon or memory. By combining a rule based system for regulars with an associative system for storing irregulars, it aims to capture both the generality and symbolic abilities of a rule system with the ability to handle arbitrary and family resemblance patterns of irregulars. The fact that regulars can exist in memory account word frequency effects and lack of total loss of regulars in neuropsychological cases. Whilst the regular ending should be applied by default in cases where memory fails, such as rare words or pseudo words. The issue of systematic regularization. This can occur in both nouns and verbs when a normally irregular form is regularized in certain contexts or compounds e.g. several Mickey Mouses versus several dormice or I was not put out versus the data were outputted (a regular verb form of the noun output). Systematic regularization is not captured directly in the model show in figure 1.5(b) but is a consequence of other allowing other grammatical rules that treat words as units rather than associative agglomerations of features.

The debate within language is far from resolved as evidenced by the exchange between Pinker and Ullman, and McClelland and Patterson in a recent issue of Trends in Cognitive Science (Pinker and Ullman, 2002b; McClelland and Patterson, 2002b,a; Pinker and Ullman, 2002a). However, for the purposes of this review, we are concerned with the definitional aspects of the debate rather than the interpretation of the data. Although, here too there are disagreements about exactly what criteria can be used to establish or disprove the symbolic or the associationist claims about rule use. This should not be too surprising because the intensity of
the debate surrounding the past tense is precisely a consequence of it’s wider theoretical implications. If aspects of language, a traditionally canonical rule-based system, can be explained without recourse to abstract symbolic manipulation then this can apply to other cognitive domains. McClelland and Patterson (2002a) want to take the successes of the connectionist models of the past tense as evidence that more generally “cognitive processes are ... graded, probabilistic, interactive, context-sensitive and domain general” (p. 465). Counting against the opposing view, namely

Advocates of symbol-manipulation suppose that there are mentally represented rules that describe relationships between variables, that those variables may be instantiated with particular instances, and that there are operations such as copying and concatenation that perform computations on variables. (p. 244, Marcus, 1998b)

In terms of defining the nature of rules themselves, McClelland and Patterson (2002a; 2002b) admit that support for genuine abstract rule use would be given by several types of evidence;

(1) Rules expected if acquisition is abrupt, that gradual acquisition is characteristic of an associative learning mechanism whereas abstract symbolic rules are more likely to be acquired in an all or nothing fashion.

(2) If the rule is applied uniformly, independent of other factors such as phonological similarity or semantic features (concreteness, familiarity, etc.).

(3) The rule process is separate from other mechanisms, thus rules could be selectively impaired in neuropsychological conditions and modularity should be expected.

(4) Regularization is a consequence of the influence of a majority of exemplars. Thus regularization should not be expected when these are in a minority.

They give the German -s plural as an example.

Interestingly, Pinker and Ullman (2002b; 2002a) contest all four of these defining characteristics of rules. They claim that during learning rules can be probabilistic and so performance improvements can be gradual rather than stepwise counting
against the first criteria. That the interaction with the associative lexicon can lead
to rule use always being seen in context which counts against the second criteria
and also, by way of compensatory mechanisms, against the third. Whilst, they
contest the specifics of the German -s plural example, as a case where neither
theory currently provides a convincing explanation of the pattern of regularities.
Instead, Pinker and Ullman claim the keys issue is whether human language is a
system that is (1) combinatorial and (2) operates on grammatical structures. As
can be seen in the past tense debate these issues are confounded by the complexity
of the real world examples. To get round this studies of the unique properties of
language are conducted using artificial grammars.

1.6.2. Artificial Grammars. Artificial grammars are systems of explicitly
constructed rules about the combination of completely artificial and meaningless
symbols. Using these rather than real language provides a vastly simplified and
constrained system whereby symbol manipulation ability can be investigated. We
will look at two aspects of language acquisition that have been investigated using
artificial grammars, infants ability to generalize in grammar learning tasks and
the performance of human and non-human primates with recursive or finite state
grammars.

Marcus et al. (1999) performed an artificial grammar learning experiment with 8-
month old infants in an attempt to show that 'algebra-like' rules are necessary to
account for the ability to generalise. Infants were habituated with a sentences in an
artificial language that had one pattern of words or another (e.g. wo wo fe or wo fe
fe, that is ABA or ABB structures.) At test using completely novel syllables (e.g. ti
ga ga or ni ni la) the infants would showing a novelty preference for sentences with
the unfamiliar grammar. This indicates that they have generalized something about
the structure of the original stimuli and by using an AAB versus ABB structure
the design ruled out the possibility that infants were just detecting the adjacent
repetition.

Several attempts were made to create connectionist models that generalize in the
same way. Seidenberg and Elman, 1999; Altmann and Dienes, 1999; Christiansen
and Curtin, 1999 all claimed to have replicated the behaviour without the need for
rules or explicit negative feedback. Marcus (1999c,a) provided a critique of these
attempts pointing out how certain modelling assumptions and implicit sources of negative feedback accounted for the results, negating their applicability to the infant learning. Further debate took place between Marcus (1999b) and McClelland and Plaut (1999). But finally, Altmann (2002) was able to produce a model that met all the criticism by providing it with prior experience of other language examples. Here exposure to the ABA grammar sat on top of an existing wider knowledge base and the network was able to generalise.

Chomsky (1957) showed that human languages have a minimal generative complexity greater than that of simple finite state machines. Rather they require a further level of computation to allow for embedding of phrases in other phrases and for the determination of long range dependencies. Using simplified artificial grammars also allows for the testing of the claim about the combinatorial nature of rule use, that can be isolate different aspects of human competence and also provides a mechanism for testing performance of non-human animals. Differences have been found between finite state grammars such as \((AB)^n\)

```
  A B A B A B
```

and a recursive or phrase structure grammar like \(A^n B^n\).

```
  A A B B
```

For example, Fitch and Hauser (2004) demonstrate a dissociation between human and non-human primates in the ability to handle recursion. Whilst Friederici et al. (2006) show that in humans the frontal operculum can detect violations of a previously learnt finite state grammar but not for a recursive grammar, whilst Broca’s area can detect violations in both types.

Hauser, Chomsky and Fitch (2002) attempt to understand what is unique about human language by comparing it to the best efforts of other animal species. They wish to identify the narrow set of features, unique and essential to faculty of language (FLN) and contrast these the broader set of prerequisites that would also be required for language (FLB), such as the ability to vocalise or hold items in working memory. They make the case that "FLN is the abstract linguistic computational system alone" (p. 1571, Hauser et al., 2002) and that recursion is the key defining feature of this system. Whether this interpretation is accepted or not, the
research with AGL shows that recursion and abstract pattern matching are two aspects of human rule use that can be appear in more abstract context. With both infants and non-human primates, the use of artificial grammars provides a means to study some of the canonically rule-like aspects of language in isolation from the supporting structure of a full adult competence with a particular language. Simple abstract rules are also used to investigate the development of executive function and evidence suggests that the prefrontal cortex (PFC) is responsible for certain types of rule use.

1.7. Rules and the Prefrontal Cortex

Within the literature on the development of executive function, young children’s capacities for cognitive control are often tested by studying their ability to apply simple abstract rules. Therefore, a consideration of this literature will provide us with some evidence about the earliest emergence of rule use in any form. Additionally, it is hypothesized that these rules are represented in the PFC and we have already seen the PFC implicated in rule use in Anderson’s ACT-R model (figure 1.4). Developmental and neuropsychological evidence (Ashby et al., 1998; Ashby and Spiering, 2004) suggests that the PFC mediates the learning of rule based categories. Whilst Miller and Cohen (2001) assign a central role to the PFC in cognitive control and executive function through the active maintenance of appropriate goals. It seems that structures in the PFC are particularly well suited to the balance the competing computational demands of requirements for applying the specific criteria of a rule and also being able to generalize it or to switch to some other rule. In this section we will outline the types of trade-offs involved in using rules and also survey some tests of executive function to determine what they can tell us about the nature of rules.

Consider for example the Dimensional Change Card Sort (DCCS) task (Frye et al., 1995; Zelazo et al., 1996). A set of cards varies on two dimensions, typically colour and shape and child is first required to sort them according to one dimension. For example, in a set of yellow and green cars and flowers, the initial rule may be colour and they are required to put all the yellow ones in one pile and the green items in another. The rule then changes to shape and children are informed that they must put all cars in one pile and flowers in another. At three years old children
tend to perseverate on whichever goal they were originally given but by five years old they can typically switch as required (Frye et al., 1995). Further experiments indicated that the younger children both understood the instructions and could verbally report the correct post-switch rule but failed to perform the correct action (Zelazo et al., 1996).

The DCCS is typical executive control task and one that illustrates the different types of demand of a rule-based task. On one hand representations must be tied closely to specific response and maintained. The children must know that colour is the important sorting dimension, responding accordingly while ignoring the similarity in shape over a number of trials. On the other hand, they must also represent this goal at a higher level in order to be successful when the sorting dimension switches to shape. Thus, there must be both a specific grounded and a general abstract representation in operation at the same time. The fact that children verbally report the correct rule but perform the wrong one could be an indication of a failure to resolve this conflict. An additional element of conflict less evident in this task comes from the need to generalise the rule on only the appropriate characteristics of the stimuli. The PFC is widely thought to be particularly well suited for these kinds of neural processes.

Bunge and Zelazo (2006) speculate that the representations in the PFC could form a hierarchy of increasing abstraction and complexity. See Figure 1.6. They tie increasingly complex rules and increasingly abstract representations to increasingly anterior portions of the PFC and ascribe particular rule types to certain PFC sub-regions. Simple stimulus-reward associations are handled by the orbitofrontal cortex, multiple conditional rules are handled by the ventrolateral and dorsolateral PFC and the rostrolateral PFC is involved in the explicit consideration of different task sets. They claim that this model gives an account of developmental changes in rule competency that is closely tied to the corresponding maturation of the cortical areas in the PFC.

Furthermore, the organisation and operation of the PFC could give insights into the algorithms of rule use. It is close to the motor cortex and so premotor areas could provide the grounded sensorimotor representations whilst the PFC also has extensive connectivity to many other brain areas allowing it to be a site where
Figure 1.6: Rules can have increasing complexity that form a hierarchy. From the simplest case where a single stimulus is associated with a reward, through sets of univalent rules where there are several alternative stimulus-response pairings, up to bivalent and higher order rules where one or more conditions determine which pairings are appropriate. More complex rules are represented more to the anterior of the PFC. Numbers refer to Brodmann areas. Diagram from Bunge and Zelazo (2006).
actions are planned and coordinated. Miller and Cohen (2001) speculate that the PFC modulates activity, it does not carry out tasks itself but helps select the appropriate response, possibly by inhibiting alternatives. Indeed this is the role of the PFC in Anderson’s ACT-R model. Moreover, the PFC is primarily responsible for maintaining representations whilst the basal ganglia is implicated in updating the choice of representation (see Ashby and Spiering, 2004).

This pattern of widespread connectivity and modulation of other processes could meet Alan Newell’s requirement for a mechanism that provides distal access to knowledge bearing structures elsewhere, most likely the inferotemporal cortex. Moreover, this could be an interactive process that accounts for learning, generalization and expertise. The existence of abstract, distal and possibly temporary representations in the PFC could aid the refinement of meaningful, semantically rich permanent representations in the inferotemporal cortex, inhibiting of inappropriate representations and allowing the diffuse network of relevant activations to be kept active whilst the appropriate associations develop between them via a process of Hebbian learning. It is known from studies of perceptually grounded category expertise that the inferotemporal cortex supports complex and abstract representations (e.g. Gauthier and Tarr, 1997; Kanwisher, 2000). It could be that in cases where the concepts are more abstract active maintenance by PFC is necessary for learning or generalization to occur. So far there are few models of how this mechanism might work However, Rougier et al. (2005) present a neural network model demonstrates flexible cognitive control without use of symbolic representations using PFC-like neural mechanisms.

We conclude this section by looking rule use in cognitive control from a different perspective, listing six criteria for inferring that rules are being used which were identified by Reese (1989) as part of a survey of rules in cognitive control.

(1) Target behaviours are regular.
(2) The development of target behaviour is discontinuous.
(3) Participants show awareness of rule use.
(4) Observed behaviour is consistent with expected behaviour

And two criteria that strengthen this requirement
(5) When there is more than one kind of expected behaviour

(6) When the rule generalizes to other tasks.

These criteria emphasize several slightly different aspect of rule-use not accounted for in the other executive control literature surveyed here. Reese’s criteria 1 and 4 remind us that rule use should be consistent over time, both within and between trials. We should expect participants to always respond in a like fashion to like stimuli. Criteria 2 goes further and speculates that learning is saltatory, either the participant can successfully utilize the rule or they cannot and the transition is abrupt. Criteria 2, 5 and 6 get at the hierarchical and abstract nature of rule use in executive control. Criteria 3 says that participants ought to be able to utilise meta-representations, whilst criteria 5 and 6 are speculations that apparently similar rules are in fact relying on the same underlying representation, analogous to Smith et al.’s criteria 7 & 8. We now move on to rule use in category learning.

1.8. Rules in Concept Formation

Category learning is defined as the ability to treat discriminably different instances in the same fashion. As such this encompasses cases where the grouping of items within the category is formed of cases that either obey some rule or set of rules and those that do not. In this section the primary focus will be on the definitions of rules. As we have seen Ashby and Spiering (2004) define rule based tasks as “those in which the categories can be learned via some explicit reasoning process” (p.104, Ashby and Spiering, 2004). Hampton (2005) goes further than Ashby or Smith et al. (section 1.5.5) in this respect and, commenting on Pothos, he proposes that “the notion of ‘rule’ in cognitive science should be restricted to those rules that can be explicitly stated by the person following the rule” (p.26, Hampton, 2005), choosing this definition precisely because this will make the notion of ‘rule’ into a empirically tangible concept. Meanwhile, Pothos define rules as ‘determined by a small subset of the relevant object properties” (p.2, Pothos, 2005). As we have remarked previously, this definition is hampered by the need to interpret what is meant by ‘small’ and by ‘relevant’. Therefore this section will look at other approaches to rules in category learning.
Early work on category learning, considered sets of necessary and sufficient conditions as one possible way of defining categories. In section 1.3.2, we have already covered why this approach does not capture the full range of categories that are possible. Here, we may consider that perhaps determining if an object possesses a single feature might be an example of rule use and that categories defined by a single feature are examples of rule based categories. However, recalling our discussion of figure 1.3 and the process of abstraction, we claimed that this process could be rule or similarity based. For example, consider the sets of all triangles, all squares, and all red objects. Membership of last is clearly requires at least one step determined by a similarity judgment concerning the ‘redness’. The set of all squares could potentially be constructed by applying particular definitional rules about squares but in practice is more likely to involve comparison to a prototype. Finally, membership of set of triangles could be determined by comparison to a space of possible types, by some affine transformation or by a rule that counts its sides. Thus, that a category is defined by a single feature is not, in itself, a way of diagnosing rule use.

Although Pothos (2005) ultimately relies on the contrast with similarity for his relative definition of rules, he does acknowledge certain properties that can be associated with rules. In particular, he accepts that rules will demonstrate certainty, systematicity, productivity and compositionality. This is also the position of Sloman and Rips (1998) and a number of other researchers, especially Fodor and Pylyshyn (1988) (see section 1.5.6) We, therefore, give brief definitions of each of these properties. For brevity we will explain these properties in relationship to categorization but the basic ideas behind them should map fairly clearly to other domains (e.g. reasoning or language)

1. **Certainty** - Concerning category membership, we should expect a rule to give us a simple mechanism of confirming or refuting category membership for any item we apply it to with a high degree of accuracy and reliability.

2. **Systematicity** - Related to certainty, a rule is systematic if it can be applied in the same way to all items in a same class.

3. **Compositionality** - Refers to the ability to combine two or more rules to produce a new category.
1.8. RULES IN CONCEPT FORMATION

(4) *Productivity* - The claim that we can combine and recombine rules without limit.

None of these properties is without difficulty. For our definition of certainty, we would need to define 'simple' in some non-controversial way. Systematicity requires a pre-existing way of determining that all items are from the same class, when in fact it might be that our rule is itself intended to determine this. A glance through any journal dealing with the philosophy of language will show that compositionality and productivity are far from being universally accepted concepts (see for example Fodor, 2001 or Szabo, 2001). However, whether these criteria can be applied with certainty could be an issue to leave to philosophers. Whilst we can use these criteria in a more heuristic way. We may take account how and why a result may be combined with others to determine what if it will have a rule like structure. For example, when considering evidence from speech perception McMurray and Gow (2005), suggest that the dimension of voice onset time can be treated as graded or discrete depending how that information is subsequently processed. In fluent speech recognition a graded representation is better for downstream processing which needs to take account of correlations with preceding and following sounds. Whilst in a phoneme categorization task this variation can be discarded as noise, leading to a discrete categorical perception.

1.8.1. Rules as simple (verbal) descriptions. As we have already seen Ashby et al. (1998) took the property of being verbalizable as their criteria for rule use, although this definition was later widened by Ashby and Spiering (2004) to anything that was learnable "by explicit reasoning processes" (p.104, ibid.). Interestingly, a somewhat different conception of rule-based categories is suggested, in passing, by Nosofsky (1986); "A rule-based approach might predict the category partitioning according to some criterion of 'economy of description' or 'simplicity of organization.'" (p.51, Nosofsky, 1986). Unfortunately, while such a criterion holds a great deal of intuitive appeal, in practice it is itself economically described by the phrase 'easier said than done', which is to say that determining a principle of simplicity is actually an extremely complex problem.

Chater (1996) describes an approach to perceptual organisation in terms of simplicity, using a definition taken from a branch of mathematics called *Kolmogorov*
complexity theory. Here the complexity/simplicity of a set of data is determined by the length of the shortest computer program that could generate the data. Chater proves that maximum likelihood and simplicity are formally equivalent, namely that the most probable hypothesis about the data provides the simplest/shortest description of it. Clearly the perceptual system will never find the absolutely simplest description of sensory data but the equivalence also applies at an approximate level. Presented as an account at the computational level (after Marr, 1982) it is perhaps more accurately be characterised as an asymptotic ideal of what the cognitive system might be capable of, but for limitations in representational power and in the ability to search amongst representations, both of which are highly non-trivial constraints on the problem.

Pothos and Chater (2002) apply this principle to a set of problems in unsupervised category learning and show that a simplicity coding can predict some category groupings. The simplicity calculation itself is of little practical use. The problem comes from the need to consider all possible partitions of N items, a numbers which grows very, very fast as N increases$^3$. However, Chater’s (2005) proof shows that a maximum likelihood approach would be equivalent and, fortunately, there are a very large number of maximum likelihood estimation methods in machine learning (Mitchell, 1997).

1.8.2. Hahn and Chater (1998). Hahn and Chater (1998) also surveyed rules and similarity. Their answer to the problem is summarized in by figure 1.7. Similarity is the name given to those processes that operate by partial matching of representations with little or no abstraction. Whilst Rules are those mechanisms that have a high degree of abstraction and require a strict match between the representations. This section will examine this proposal in greater detail, covering the reasons why Hahn and Chater propose it, how they treat representations and whether their proposed dimensions of abstraction and partial-strict matching are meaningful and empirically useful.

Hahn and Chater’s proposal starts with a careful consideration of exactly what is meant by ‘representation’. They begin with the observation that it is not the way that an object is represented that determines whether it is classified using rules (or

---

$^3$15 items permit just over a $10^9$ partitions while just 75 items can be partitioned in about $10^{80}$ ways, a number comparable to the number of particles in the known universe.
Figure 1.7: The relationship between rules and similarity in categorization as envisaged by Hahn and Chater (1998). Similarity processes are characterized by a combination of partial matching of items and little or no abstraction. Rule based mechanisms are highly abstract and rely on a strict match for items that they apply to. The bottom right quadrant refers to memory processes whilst it is unclear what cognitive interpretation could be given to the remaining areas.

similarity). One object can be classified in many different ways using more or less of its properties, as Barsalou’s work on \textit{ad hoc} categories shows (Barsalou, 1983). In Hahn and Chater’s scheme, rules and similarity differ in how new examples are matched to existing knowledge. In the case of rules, the match is \textit{abstract} and \textit{strict}; only a subset of the properties must be matched and each property has to be matched with a high degree of certainty.

First, note that this is a very different definition of abstraction to the one given above (section 1.4.2.2). In that case abstraction was a process that extracted a single property or feature needed for the application of the rule, we claimed that this could be either a rule or a similarity process. In Hahn and Chater’s framework, abstraction refers to the selection of a few key properties from a larger set, where the properties themselves are already given. In this way they have failed to address the prior problem of how those properties are obtained. This is a real problem. One of the main aims of the category learning literature is to explain how new categories can be acquired, central to this is a need to assimilate new basic properties.
To see how this is problematic consider the famous 'duck test': "If it walks like a duck and quacks like a duck then it must be a duck." It meets one of Hahn and Chater’s criteria for a rule use in that it uses a subset of the available properties of ducks for duck classification. However, walking like a duck and quacking like a duck are both examples of typical similarity based judgments and by Hahn and Chater’s criteria can only result in a partial match. Putting this example in the problematic top-left quadrant of their classification. One reaction could be to use their definition to get around the problem mentioned in section 1.4.2.2 above by not allowing that property matching itself could be a similarity process. However, Hahn and Chater do not appear to require this and even give an equivalent of the duck test, 'the dog rule' ("if something barks and is furry, then it is a dog", p. 201, ibid) as their paradigmatic example of a rule, yet both individual properties appear to be similarity based. Unless further assumptions are added about the ability to strictly match single features or dimensions, this makes it difficult to carry Hahn and Chater’s formulation any further.

1.9. Rules in Animal Cognition

Whilst animals show a wide range of very impressive behaviours and learning abilities, there has only been very limited success in demonstrating genuine abstract rule use in animal cognition. Nonetheless, there are two ways in which a consideration of animal cognition can inform a study of human rule use. First and foremost because it is important to be aware of the power of associative learning and instrumental conditioning. Which will help guard against the positing of rule use when simpler mechanisms can account for the same behaviour. This is a general principle in behavioural research first articulated in the comparative learning literature by Lloyd Morgan:

\[
\text{In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of one which stands lower in the psychological scale} \\
(p.53, Morgan, 1894)
\]
Secondly, in those few cases where animals do show rule governed behaviour, the conditions are likely restricted to very simple cases. In this section we look at what does and what does not constitute rule use in animals.

One of the first questions we should address is why classical stimulus-response and operant conditioning are not examples of rule following. We deal only with classical conditioning but similar considerations apply to operant conditioning. The Rescorla-Wagner model of Pavlovian conditioning (Rescorla and Wagner, 1972) is a simple model that accounts for many of the features of basic animal learning. The basic idea is that learning is proportional to the amount of 'surprise' the animal experiences on each pairing of conditioned stimulus (CS) and unconditioned stimulus (US). The model accounts for differential learning rates, patterns of extinction, blocking effects and the conditioned inhibition. Whilst Pavlovian conditioning can link a (relatively) arbitrary CS to a response this is far short of rule learning. The CS is always a single perceptual stimulus or a very narrow perceptually similar class of stimuli, thus falling short of the abstraction property widely ascribed to rules. Additionally, responding is not all or nothing. Whilst learning does not show any abrupt transitions and does not transfer to novel or analogous situations.

Despite this there are some instances where animals could be said to show rule-like behaviors. We have already encountered the case of recursion in section 1.6.2. One example is literature on the ability of various animals to learn the concepts of same and different. Pearce (1997) reports that with sufficient training chimpanzees, rhesus monkeys, dolphins, sea lions and crows have all been shown to generalize based on this type of relation but that his own partial success in teaching pigeons the principle might in fact be attributable to them exhaustively learning all possible configurations (Pearce, 1988). Young and Wasserman (2002) showed that pigeons found same and different discriminations easier with larger arrays of items. But this finding suggests that rather than using an abstract rule pigeons are learning to respond to the 'visual entropy' of the array. We return to these results and the comparative dimension of the question in Chapter 3 below.
Historically rule use in adults was often assessed on the basis of verbal reports. This is clearly an insufficient basis for concluding that this in fact how participants are solving a particular problem. Firstly, it is possible that the participants are using different criteria to those which they report. Indeed, it may be that the requirement to provide a verbal report itself forces the participant in an experiment to simplify and formalize the more intuitive process that they actually used in a task.

An alternative method of assessing reasoning is to compare performance to some pre-determined possible models of response. For example, one might expect one group who respond randomly, another group who follow a partially correct rule and a final group who respond correctly throughout. Individuals can then be assigned to which ever group best matches their performance. This is the basic idea behind Siegler’s rule assessment methodology (Siegler, 1981) and Thompson’s rule-testing framework (Thompson, 1994; Thompson and Markson, 1998). These approaches run into two problems. How does the researcher determine what rule sets are possible. Secondly, they require a principled statistical way of dealing with individual differences, especially individual error rates which can be high in developmental studies. Similarly, if Ashby’s hypothesis is right and all category learning tasks involve a competition between rule- and similarity-based processes then this type of analysis is useful as data will be intrinsically noisy.

Imagine that in a Dimensional Card Sorting Task (DCST) there are classes of participants; children who are guessing throughout the experiment, children who succeed at the first task but perseverate after the dimension change and children who successfully switch. However, both these latter two groups occasionally make incorrect responses. An experimenter will just have the response data and will in general have no external way of knowing which pattern of responding a particular child. The best that one can hope for is to assign the participants into the respective classes in a way that maximises the likelihood that one has the correct distribution. Latent Class Analysis (LCA) and Finite Mixture Modelling are two closely related methods that do this.
1.10.1. **Latent Class Analysis and Finite Mixture Models.** Schmittmann et al. (2006) modelled a slightly simpler task where the participants, aged between 4 and 20 years old, were required to determine (based on feedback about their guesses) which of two uni-dimensional rules was being used to classify items. Items were either black or white and either triangles and squares. They tested several finite mixture models and found that a two mode model provided the best (statistically most likely) explanation for the data. Some participants, particularly those older ones, used a hypothesis testing approach where learning was sudden, showing systematic testing of responses and high accuracy once the correct hypothesis had been found. Whilst other participants, particularly the younger children, learnt in a slow and gradual fashion. This type of analysis is particularly useful in tasks with much younger participants, where the error rates will be large and the strategies are less in accord with our rational expectations.

Latent class analysis operates analogously. For example, Jansen and van der Maas (1997) used LCA to reevaluate Siegler’s (1981) rule assessment methodology of the balance scale problem and to assess the fit of a PDP model of the task (McClelland, 1989). In doing so they were interested in seeing how useful a consistency criterion was when the observed data was likely to be noisy. The classes are sets of probabilities of correct responses on given balance-scale problem types, and as such certain constraints can be added according to rules that might be expected (for example, one group might guess on every type of problem while another might be correct on all problems with equal weights on either side of the fulcrum). The latent classes analysis estimates both the number of different strategies and the proportion of participants using each strategy. Jansen and van der Maas (1997) found evidence of rule use in the empirical data but not in the PDP model data. This demonstrates the ability of LCA to find rule-like responding in noise and equally not to mistake noise for rule-like responding. As such, it is a very statistically well-principled way of judging when rules are being used.

1.10.2. **Bayesian Rules.** Up to this point, even these techniques treat rules in a ‘traditional’ form. Although LCA and FMM add the means to deal with uncertainty and error in use of a given rule (the unconditional probabilities in the LCA model) and also with heterogeneous sets of rules between individuals (the
conditional probabilities). These methods still treat rules in a classical fashion, each rule is treated as an independent logical operator acting on a objectively uncontroversial dataset. It is a frequentist approach. The Bayesian theory claims this is an incorrect interpretation, that probability should be viewed as subjectivist or Bayesian. This approach inextricably links the interpretation of data to set of expectations about that data. This can be seen from equation 1.2 which is a simple expression of Bayes’ Theorem.

\[
P(Hypothesis \mid Data) = \frac{P(Data \mid Hypothesis) P(Hypothesis)}{P(Data)}
\]

To translate into the language of rule use it says that arriving at best interpretation of a rule when given a set of data (maximizing the left hand side) is equivalent to defining a rule that makes the most plausible interpretation of the most data with the fewest assumptions (maximising the quotient and minimizing the denominator on the right hand side.) Or to phrase it slightly differently the chance that a particular outcome happened depends on what we know about this particular occasion and our prior experience of events like this. It is this second part of the equation the Bayes’ formula emphasises, the importance of prior probabilities. This is made clearer by an example;

Imagine that you are calling your friend Thomas on his mobile phone. You know he is on holiday so one of the first questions you ask him is about the weather. Rather than answering you directly he merely says "Well, we’ve got the umbrella up," before turning the conversation back the far interesting topic of developmental psychology. Stuck in your office, your own interest in the weather on Thomas’s holiday is little more than academic. But to answer the question, it is advisable that you use equation 1.3.

\[
P(Rain \mid Umbrella) = \frac{P(Umbrella \mid Rain) P(Rain)}{P(Umbrella)}
\]

The left hand side of your equation is the familiar (if misleading) conditional probability of there being rain if the umbrella is open. But this is better understood by
considering the right-hand side. For simplicity’s sake, we may assume that Thomas
is a sensible fellow who would use his umbrella if it were raining, so the first term
on the right is pretty close to one. The other terms involve our assumptions about
the prior probabilities. The likelihood that it is raining where Thomas is on holiday
and the likelihood of there being open umbrellas. Since you know that Thomas is
on a beach holiday in the Mediterranean you know that the absolute chance of rain
is very low and that there will be quite a lot of people with open beach umbrel-
las. Using this information you decide it is quite unlikely that Thomas’s original
comment implies that it is raining and you start to think about booking your own
holiday. If Thomas had been on a walking holiday in the Scottish Highlands the
outcome would have been quite different.

This illustration hopefully gives a clearer sense of how Bayes principle applies but
this is not to say that people are always Bayesians. Evidence suggests that peo-
ple are bad at reasoning in a probabilistic fashion when presented with explicit
judgments about probability and likelihood (Tversky and Kahneman, 1974; Kah-
neman and Tversky, 2000). Although performance can be dramatically improved by
emphasising frequencies rather than probabilities (Gigerenzer and Hoffrage, 1995).
Conversely, however, people appear to reason by Bayesian principle in circum-
stances where they might be expected to be logical. In their book Bayesian Ratio-
nality, Oaksford and Chater (2007) apply the Bayesian approach to the canonically
logical and rule based fields of syllogistics and conditional logic. They argue against
the traditional view that when people make ‘logical errors’ in reasoning tasks it is
due to the imperfections of the human reasoning process. Rather, that when anal-
ysed according to a Bayesian framework, there is a sense in which the choices
individual make is optimal. Thus, people appear to use probabilistic rules rather
than strict logical quantifiers.

1.11. Some Rules about Rules

We have surveyed a large number of opinions about what rules are in cognition and
by what criteria we may know if they are being used. In this section, we attempt to
summarise and synthesise the work surveyed above with the aim of producing a set
of defining characteristics for judging whether behaviour in concept learning to be
rule-based or not. Or failing that to provide the strongest statistical expectation
that a pattern of results is due to some underlying rule. Additionally, we shall consider how rules might appear from a developmental perspective.

Recalling the two stage model of abstraction and application and referring to Hahn and Chater’s (1998) definition of rules as strict matching plus abstraction, one of the first steps might potentially be to establish that some degree of abstraction has occurred. This is also part of the definition in Pothos (2005) of a “small subset” of available properties. This would ideally be done by explicitly identifying the dimensions or features that were relevant. It also implies that there will be other features which are ignored. The strict matching criteria or a high degree of certainty about rule application is considered critical by several authors. This is also related to the criteria about how rules generalize. If rules are used then one might expect equal performance with novel materials, or when comparing concrete to abstract performance.

While several authors (Reese, 1989; McClelland and Patterson, 2002a) speculate that learning of rules will be abrupt, is there any principled theoretical reason for this or any consequences if it were found to be the case? For example, would it imply that rules are acquired by a rational process of hypothesis testing? Would a Bayesian analysis have anything to offer in this respect? Alternatively, do the explicit models built offer the best starting point for predicting how old rules will be used in new situations, such as within the ACT-R framework.

Finally, the idea that rules are equivalent to just those descriptions that are verbalizable or can be explicitly learned (Ashby and Spiering, 2004) or even are just those rules that a participant verbally reports (Hampton, 2005) appears not to capture the full range of possible rules, but what direct test could establish this? Meanwhile, the idea that rules capture some 'simplicity of organization.' (p.51, Nosofsky, 1986) has intuitive appeal and perhaps the simplicity approach of Pothos and Chater (2002) could be adapted to give a meaningful definition of a rule. While the simplicity concept does not lead us directly to optimality because of the combinatorial explosion of possibilities, the simplicity measure can be used to compare the relative simplicity of candidate rules. In summary, there is a broad consensus about many of the features of rules and their use. However, there has been a noticeable lack of a developmental perspective in most of the work surveyed here.
1.11. Development of rules. There are three contributions that a developmental stance offers to our consideration of rules. Firstly, it can take us back to first principles, looking for the absolutely simplest instances of rule use and how they might be identified. Secondly, it raises the developmental question of what rule representations are possible at what age and by what mechanisms can development of more sophisticated abilities be attributed to. Finally, a development perspective focuses on how learning might be happening and allows the possibility that different ages will learn the same rule in different fashions.

Considering the first of these issues various subsidiary questions arise from any answer we give. If hypothetically we decide that the rule ‘bigger than’ was an example of one of the simplest possible rules, then we would be interested in discovering if, once acquired, it was applied consistently across a range of context and in the presence of various distractors. We might also ask what would half a rule look like? Or is it possible to know what a rule is but not know how to use it? The results of Zelazo et al. (1996) showed that children develop the ability to represent a dimension change rule but were unable to utilise this information correctly.

Furthermore, is the development of rules to be equated with the ability to utilise increasingly complex representations or do other things develop too, such as an increasing accuracy, improving reaction times, ability to ignore distractors? Sitting on top of these considerations, is the complication that experimental data from imperfect learners will be noisy and principled statistical methods will be required to determine if particular rules are being used. Similar statistical considerations apply to any investigation of the process of learning. How should abrupt learning be characterised? Might abrupt changes in individuals performance be masked in group data?

At the risk of complicating things even further, figure 1.8 attempts to demonstrate several possible ways that a rule might develop. If we consider the final learned state for the hypothetical rule to be a step function along a single dimension, there are at least four ways that this might be acquired. In case (a) the rule criterion or boundary might be known from the start but errors are introduced in responding. In case (b) the rule may be perfectly applied at the extremes and there may be an expectation of an abrupt criterion but there is uncertainty about where that
1.12. SOME PROBLEMS WITH SIMILARITY-BASED CATEGORIES

At first glance, some measure of the similarity between items seems like an effective mechanism for building categories. And indeed Rosch (1975) proposes that a good category structure will maximise the within category similarity and maximise the between category difference. But there are various theoretical objections to the
idea that similarity alone is sufficient grounds for establishing category structure and several empirical findings that suggest that judgments of category membership and ratings of item similarity do not always coincide.

1.12.1. Goodman’s Paradox. To establish how similar two items are, it seems like one might compare their properties to see how they measure up to each other. Thus, an alligator and a crocodile are found to be creatures of comparable size, colour and appearance that live in comparable habitats. Whereas, if we are asked, as the Mad Hatter asked Alice, in Lewis Carroll’s Alice in Wonderland, “Why is a raven like a writing desk?”, we might pause for a second or two before concluding that they are completely unlike each other. But the philosopher Nelson Goodman points out any two items are alike in any number of ways; ravens and writing desks each weigh less than a tonne, have more than one leg, and are a similar distance from the sun, etc. This is known as Goodman’s paradox (Goodman, 1955, 1972) and it shares a great deal in common with Wittgenstein’s observation about the under-determined nature of rules (Section 1.5.1).

One possible resolution is to suggest that we attend to just those properties that are psychologically relevant or salient. It is certainly true that there are a large number of constraints imposed by our perceptual system. Unlike bees, we do not notice any similarities or differences the ultraviolet colouration of flowers. Conversely, as Goldstone (1994) points out, “It is difficult not to notice the similarity between a 400 Hz tone and a 402 Hz tone.” (p. 136, Goldstone, 1994, emphasis in original). Nevertheless, Goodman’s argument is that similarity alone is insufficient to build categories because we must always specify the dimensions with respect to which the members of a category are similar or different. A raven and an ostrich are both feathered bipeds but there the similarity ends, while an apple, a tomato and a cricket ball are all red, shiny spheroids of approximately equivalent size and weight but only the first two are fruits. Goodman claims that the similarity of members of the same category must be accompanied by a specification of which dimensions are relevant and therefore it is this step which constructs the category rather than any \textit{a priori} measure of similarity.

A slightly more principled argument against Goodman’s position may proceed along Bayesian lines suggesting that the relevant dimensions are those that have proved
diagnostic in the past. So for example, the feature of being 93 million miles from the sun is ignored because it does not discriminate between any of the objects we typically encounter. Kemp et al. (2007) show that a hierarchical Bayesian model can account for young children’s acquisition of the shape bias in word learning (the expectation that similarly shaped objects are of the same kind). However, Bayesian models can often build assumptions into their prior probabilities. For example, Kemp et al.’s (2007) model was provided with category labelling information and only has to determine which type of properties were predictive of those categories. Thus it is not a sufficient model for completely bootstrapping categorization and does not as it stands provide an escape from Goodman’s paradox.

Likewise, Tenenbaum (2000) and Tenenbaum and Griffiths (2002) emphasise that it is understanding the power of generalisation rather than building a theory of category structure that motivates their Bayesian approach. This may be important to understand how people draw likely conclusions on the basis of very limited evidence by adopting the most probable hypothesis about those data. But it does not account for where those hypotheses originally come from. Tenenbaum and Griffiths (2002) acknowledge this and speculate that the hypothesis space is built by a combination of evolutionary constraints and prior learning. Furthermore, they argue that the notion of similarity that arises out such an approach is potentially able to reconcile the set theoretic notion of similarity (Tversky, 1977) and the geometric conception of similarity as a function of distance in a psychological metric space (Shepard, 1987). Thus, in principle we cannot escape Goodman’s Paradox but in practice similarity turns out to be a useful guide to category membership.

1.12.2. Theory and goal dependent categories. Similarity seems less relevant in the construction and use of categories in those instances where categories appear to be based on wider knowledge about the world, such as when category is defined in relation to some goal or when some theory guides thinking about members of a category. Barsalou (1983) investigated individuals responses to ad hoc categories such as ‘things to take from a burning building’ while Barsalou (1985) considered goal-derived categories such as ‘things to eat on diet’. In both cases he found that there was a high degree of agreement among participants as to what would constitute good and bad members of such categories but there is very little
otherwise in common between the individual items. Thus similarity between item
does not offer a useful predictor of other members of the class.

Meanwhile, even in childhood, theories can determine how category members are
treated. Carey (1985) and Gelman and Markman (1986) both found that children’s
nascent biological theories would lead them to make inductive generalisations about
hidden properties that fitted better with their knowledge of those categories rather
than on the basis of visual similarity. For example, the four year old children in
Carey’s study were far more likely to extend a property (e.g. has a spleen) from
people to aardvarks (76%) than from dogs to aardvarks (29%) despite the greater
similarity of dogs and aardvarks. While Keil (1989) finds that with increasing age
children become more committed to the idea that an animal which looks exactly
like a raccoon but with skunk parents and the internal organs of a skunk really is a
skunk. Gelman uses results of this kind to advance the theory that young children
have a strong essentialist bias (Gelman, 2003). While this seems to weaken the case
for similarity, Goldstone (1994) argues that similarity need not be a unitary phe-
nomenon, and that therefore an animal may be more visually similar to a raccoon
but have more overall similarity to a skunk. Similarity is by this account seen as
flexible rather than as weak.

1.12.3. Context Effects. In fact, another set of results suggest that simi-
larity needs to be flexible because different tasks have different demands and the
context in which a similarity judgment is applied to alters the assessment of similar-
ity. For example, in Sjöberg (1972) a falcon and a chicken were judged more similar
if the set of exemplars was \{falcon, chicken, robin, wasp\} than if it was \{falcon,
chicken, robin, sparrow\}. Or to take an example from Tversky and Gati (1978),
Cuba and North Korea may be judged to be similar when the political dimension
is emphasised but so may Cuba and Jamaica when geographic features are empha-
sised. This leads us into a consideration of one successful model of categorisation
that evolved from considerations about context.

Medin and Schaffer (1978) realised that context effects are important benefit to
categorisation if categories are defined in terms of their individual exemplars and
a weighting of the importance of particular dimension they were being compared
on. The differences on relevant dimensions become more important. Furthermore,
the similarity on each the dimensions is multiplied together so that matches and
mismatches both contribute the final similarity rating. Category similarity is then
determined comparing new item to every possible category member and adding up
the similarity scores and finally a new item is assigned to the category it is most
similar to. This model has now been largely superseded by a more general model.

1.13. The Generalized Context Model (GCM)

In this section, we describe the features of the Generalised Context Model (GCM,
Nosofsky,1986; 1991) and give some sense of how it works. By studying the explicit
implementation of a similarity model we can see the assumptions that go into it and
get some sense of how different elements of such a model interact. As Pothos (2005)
mentions the prototype versus similarity debate can largely be ignored, primarily
due to the success of the GCM, which is an exemplar model but which captures
many prototype effects.

The GCM starts with a geometric notion of distance between individual items. This
assumes that there are scalar dimensions along which objects are commensurable.
(As an alternative consider Hahn et al., 2003 where objects are compared in terms
of the steps required to transform one into the other.) A further question arises
as to what dimensions are used, if an obvious set of physical and/or psychological
dimensions are evident they may used. Alternatively these may be derived from
a multi-dimensional scaling (MDS) of a large set of pairwise similarity ratings. In
this case the dimensions found may or may not correspond to actual meaningful
psychological dimensions (and although this is preferable it is not essential for the
model to work.) Equation 1.4 quantifies the distance measure.

\[ d_{ij} = c \left[ \sum_{k=1}^{N} w_k |x_{ik} - x_{jk}|^r \right]^{\frac{1}{r}} \]

It says that the distance \( d_{ij} \) between two items \( i \) and \( j \) depends on the weighted
metric of their difference along each dimension \( k \) of the \( N \) possible dimensions. The
power \( r \) determines the metric, \( r=1 \) gives the 'city-block' metric and \( r=2 \) gives the
'Euclidean' metric. Other values are possible but \( r=2 \) is the most commonly used
1.13. THE GENERALIZED CONTEXT MODEL (GCM)

(based on the best-fitting in MDS-choice models). The weighting parameters, \( w_k \) are free parameters in the model determined by the data.

The distance between any two items \( i \) and \( j \) is used to derive a value for the similarity \( s_{ij} \).

\[
(1.5) \quad s_{ij} = e^{-c.d_{ij}}
\]

Similarity here refers to the psychological notion based from participants similarity judgments. The choice of Equation 1.5 is based on the original idea of Shepard (1957; 1987) that similarity judgments do not have a linear relationship with psychological distance. Similarity decreases with the exponential of the distance, following either a straight exponential decay \( p=1 \) or a Gaussian decay, \( p =2 \). The scaling parameter \( c \) is freely fitted to the data.

Next GCM requires a decision rule by which a new exemplar can be categorised based on its relative similarity to all other items in that category.

\[
(1.6) \quad P(J \mid i) = \frac{\left( \sum_{j \in J} s_{ij} \right)^\gamma}{\sum_{K} \left( \sum_{k \in K} s_{ik} \right)^\gamma}
\]

Equation 1.6 states that the probability that item \( i \) is categorised as a member of category \( J \) is found by summing its similarity to every member of \( J \) and dividing this by the sum of its similarities to each other possible category. It is this step which makes GCM an exemplar model, a novel object \( i \) is classified using a function that takes into account its similarity to every other exemplar that the model knows about. Notice also that \( \sum_{K} P(K \mid i) = 1 \), so an exemplar is always assigned to an existing category. Another way of stating this is to say that GCM is a static model not a learning model. However, GCM is a probabilistic model returning a probability that item \( i \) is categorised as a member of category \( J \) rather than an absolute category criterion. This can adapted by changing the value of the parameter \( \gamma \). If \( \gamma=1 \) response are probabilistic with large values of \( \gamma \) the response
become more deterministic, $\gamma$ was introduced into the model by Ashby and Maddox (1993).

GCM is a highly successful model that fits a wide range of adult data. However, it has a large number of free parameters and this leads to criticism that it is too unconstrained (Smith and Minda, 2000) although these points are contested (Nosofsky, 2000). Nevertheless, GCM provides a good starting point for an example of a model similarity model of categorisation. Having surveyed the rules and similarity debate within the adult category learning literature, we now consider why it is appropriate to study category learning in infancy and survey the methods available in this field and some of the existing results and theories.

1.14. Why Study Category Learning in Early Infancy?

The majority of the research presented below will concentrate fairly specifically on the process of category learning in early infancy. Moreover, we will be particularly concerned with instances where infants are in the process of learning new categories, rather than the development and application of general knowledge about the world. There are a number of reasons for such a narrow focus.

Firstly, we are interested in investigating the processes of category acquisition. In most real world cases this will interact with prior knowledge and experience but these are very difficult to assess or control for in a systematic fashion. For example, when investigating a young infants’ knowledge of animate versus inanimate or global versus basic distinctions in category hierarchies we cannot know how much exposure infants have had to the relevant categories before coming into the experiment. For this reason, we would like to concentrate on infants’ learning of categories of artificial stimuli which they have had no prior experience. And on cases where category groupings are to some extent arbitrary, providing control that one set of infants can learn one grouping whilst another learns an independent grouping of the same items. In practice, it is not always possible to manipulate experimental categories in this way. But in such cases careful consideration must be given to the extent to which learning has taken place prior to the infant coming into the laboratory.
Secondly, for related reasons we will primarily focus on categorization in pre-linguistic infants. The interaction between concept formation and language acquisition is likely to be highly complex and a bi-directional process. The mapping of word meanings and object labels onto real world categories and classes of objects will depend on the development of language, world knowledge and categorization processes. As such any theoretical or experimental approach to categorization in this context is will have to take into account a very broad range of considerations from several cognitive domains. As such the specific conclusions about category learning will be hard to isolate. For an outline of one approach to studying the links between early object naming and concept formation see Waxman (2003).

Finally, the motivation behind this review and research is to be able to relate developmental findings to the adult literature on category learning. Category learning and category expertise are often treated separately in adult literature. Furthermore, within the category learning literature there is considerable debate about whether there are several mechanisms for category learning. Ashby and colleagues (e.g. Ashby and Spiering, 2004; Ashby et al., 1998) contend that perceptual prototypes, more abstract similarity judgements and rule-like categorization are all mediated by different brain mechanisms. Whilst Pothos (2005) suggests that the observed differences between rule-like and similarity-like categorization are due the number of dimensions that are relevant to the category groupings. The infant literature will be reviewed with an eye on how it might inform and be informed by this debate.

This review will outline the experimental methodologies that are used to assess infants’ categorization abilities. It will briefly address their strengths and weaknesses. We will then review experimental findings and the theoretical frameworks that have been built using these techniques. In addition, it will look at some computational models of early category learning and the development of categorization in the first few years of life. We include a brief section outlining how rule use and statistical learning can be thought of as a form of categorization whereby varying exemplars are all treated equivalently according to some rule or regularity. We review developmental literature in this field. To conclude, we will highlight the unresolved issues in infant categorization and how they relate to broader questions about category learning in mature individuals.
1.15. Methods for Investigating Categorisation in Infancy

The methods used to assess infants’ categorization abilities vary with age as perceptual and sensorimotor abilities rapidly develop. Some techniques are similar to those used in other areas of infancy research, whilst others are specifically developed to assess categorization skills. Several techniques make use of infants’ well known novelty preference and tendency to habituate to a stimulus or set of stimuli on repeated exposure (Fantz, 1964). Infants are habituated to items from within a category and tested for a preference for items from a novel related or unrelated category. However, specific care needs to be taken to ensure that habituation is to the category or concept. That there is not some simpler explanation for the recovery in looking times (e.g. great perceptual variation in the test items). Two other techniques; conditioned leg kick and sequential touching will be described. They are designed for younger and older infants respectively. These are both tasks that require the active engagement of the infant, but neither is without difficulties when assessing category learning as will be discussed. For a more extensive review see Mareschal and Quinn (2001).

1.15.1. Novelty Preference.

1.15.1.1. Visual Preference. Category learning in infancy is often assessed using a visual preference paradigm (Fagan, 1970). Usually a fixed training schedule is used, showing pairs of pictures of category examples presented side by side for a fixed number of trials of each of fixed length of time. The test consists of two presentations of comparison items counter-balanced so that each one appears on left and right with equal probability. The experimenter records the length of time the infant spends looking at each stimulus.

There are two possible designs and comparisons. Either a novel category exemplar compared with a novel non-category member when a preference for the non-category item would be expected if categorization has occurred. For example, Quinn et al. (1993) showed that infants prefer to look at pictures of cats after being shown multiple pictures of dogs. Alternatively, the test stimuli might be a novel prototype for the category tested together with a previously seen exemplar. Here the infant is expected to prefer the previously encountered exemplar to the novel prototype.
For example, Bomba and Siqueland (1983) exposed 4 month old infants to distorted geometric dot patterns as in Posner and Keele (1968). On test, the infants treated the prototype as more familiar. See section 1.16.1 below. In both cases, the experimenter must control for any prior preferences among the test stimuli. Furthermore, infants sometimes show a familiarity preference which could confound the data. This is more likely with short exposures or complex stimuli making interpretations of these studies problematic.

1.15.1.2. Habituation/Dishabituation. A closely related technique uses a modified version of habituation/dishabituation paradigm (Fantz, 1964; Reznick and Kagan, 1983). Infants are exposed to visual representations from a given category. The length of time they look at each picture is recorded and they are shown a random sequence of exemplars until either a pre-determined maximum number of trials is reached or some habituation criteria is attained (e.g. average looking time in last four trials is half of average of first four). By allowing the infants own attention to determine their exposure to the stimuli, this technique guards against the possibility of incomplete encoding or a familiarity preference. One further advantage of this technique is that it can be used even with very young infants (2 months) and so can provide a standard method to assess the development of categorization behaviour over the first year.

To test that categorization has occurred, three test trials are given immediately after habituation. In each trial the infants’ looking time is the dependent measure. The first item is a novel exemplar from the category to be learned. It is expected that looking time for this item will not differ significantly from the last habituation trials. The next test trial will feature a novel item from a contrast class. For example, if infants had be habituated with pictures of cats, the contrast item might be a picture of a dog. If infants are aware of the difference then they would be expected to dishabituate and look longer at this item. However, the null finding could be a failure of attention rather than failure to categorize so a third and final test trial is included featuring a very different stimulus (e.g. a truck) dishabituation to this item indicates the infants are still attending to the experimental set up.

For example, Cohen and Strauss (1979) used this paradigm to study infants’ categorization of faces. Infants were habituated to pictures of female faces in three
separate conditions. Infants in condition 1 were repeatedly shown the same face always with the same fixed expression and orientation (e.g. smiling and looking to left) and were then tested with a photo of this person with a new (neutral) expression and novel face with a neutral expression. Infants at 4, 6 and 8 months old all dishabituated to both test stimuli, indicating that all ages could discriminate facial expression (recall that categorization requires both discrimination between exemplars and generalization across the category). In condition 2, three further groups of infants were habituated to four photos of the same person displaying four different expressions and were again tested on same face neutral and a novel neutral condition. Here the 6 and 8 month olds did not dishabituate to the same face, indicating that they had formed a general category for that individual constant across expressions. In condition 3, three final groups of infants were habituated to four different female faces and were tested on a previously seen example with new neutral expression and novel face. Only the 8 month olds did not dishabituate to either face, suggesting they had formed a general category of female faces. However, this conclusion is based on a null finding and so required further corroboration (Quinn et al., 2002).

1.15.1.3. Object Examining. Object examining is a manual equivalent to the habituation/dishabituation technique (Oakes et al., 1991). As such it is only suitable for testing infants over about 7 months when they have developed sufficient dexterity to manipulate the objects. Whilst infants over about 13 months are generally too mobile to engage for sufficient length of time on such a task.

Infants are seated in a high chair and given a sequence of toys to play with. Each trial is usually of fixed length (typically 30s.) On each trial the experimenter places a toy on the tray in front of the infant and draws the infants attention to it. If the infant drops or knocks the item the experimenter quickly replaces it but otherwise does not engage the infant. Initial presentation normally consists of repeated exposure to just a few exemplars. Typically three or four objects will be each be presented separately several times. The order of presentation being randomized.

Following the same logic as the visual habituation studies, test phase consists of the presentation of three further objects, a novel category member, a novel non-
category member and a completely unrelated object. Sessions are videoed and the subsequently both visual and manual attention are scored for, using a criteria of ‘active examination’. Thus banging an object on the tray does not indicate that an infant is attending to the particular features of that object. Scoring by multiple raters blind to the condition controls for the subjective nature of this measure. However since infants are always tested with toys and models of real world items prior knowledge will be expected to interact with any on-line category learning that may take place.

It is possible that object examining is not as sensitive to category detection as visual preference paradigms. Younger and Furrer (2003) showed that nine month old infants could detect a basic level category distinction (dogs vs. horses) using a visual familiarization technique but not with a standard object examining equivalent when they showed no novelty preference. However, by modifying the object examining technique to use paired-preference presentation Younger and Furrer were able to get the infants to discriminate these categories. This suggests that different methodologies will tap into different aspects of infants category learning abilities and care must be taken about the generality of conclusions from a given experiment.

1.15.2. Operant Conditioning.

1.15.2.1. Conditioned Leg Kick. The conditioned leg-kick paradigm was developed to test memory and learning in very young infants (3-6 months old) and it also provides an opportunity to test retention of learned categories over longer time scales (from 24 hours to a couple of weeks). It has been used extensively by Rovee-Collier and colleagues (e.g.Rovee-Collier, 1983; Hayne et al., 1987; Greco et al., 1990). Infants are tested lying on their backs in their own cribs at home but at a time when then are likely to be alert. A mobile hung with some salient test stimuli is positioned over the crib and a ribbon is attached to the infants’ foot. Initially this is tied to a second empty mobile out of sight of the infant and their spontaneous base rate of kicking observed over a several minute period. The ribbon is then moved to a mobile above the crib. Kicking now causes this visible and exciting mobile to move and so infants increase their kicking rate. This phase is known as conjugate reinforcement and typically lasts 9 to 15 minutes. In a final
test phase the ribbon is moved back to the empty mobile but if the infant has learnt the association, his/her kick rate will be increased above the base line.

The category learning is investigated by giving the infant several sessions over the course of a few days each with a different exemplar mobile each time and other sessions with an unconnected mobile with a different category. For example, in Hayne et al. (1987) infants might be reinforced when they saw yellow blocks with the character 'a' in different fonts but not for yellow blocks with different '2's. When tested 24 hours later they transferred their conditioning to a novel instance of an 'a' but not a novel '2' or blank yellow blocks. Some retention was also found a week later but these sessions required a small ‘reminder’ session of where they are exposed to the mobiles and the experimenter while the experimenter moved them.

This technique is effective for demonstrating that memory for the conditioned stimuli and categories can be learned over and retained for relatively long periods (Greco et al., 1990). However, it is of limited value for investigating category learning. Firstly, it is an impractical and time consuming procedure involving multiple visits to the infants’ home which results in small sample sizes. Secondly, by typically having only around three different training stimuli, it is very restricted on the complexity of category grouping it can test. Finally, because learning takes place over the sessions and due to such other factors such a fatigue or general variability in alertness the data can be hard to interpret.

1.15.2.2. *Anticipatory Eye Movements.* A new and so far largely unproven technique is described in McMurray and Aslin (2004). It uses tracking of anticipatory eye movements when infants are presented with a conflicting category choice. Experiment 2 of McMurray and Aslin (2004) illustrates the procedure. An infant is given an fixed number (typically 30) of familiarization/learning trials where an animated looming circle with narrow horizontal stripes predicts an interesting image appearing on the left of the screen, whilst one with wide vertical stripes predicts something appearing on the right. In the next phase, infants are presented with stimuli that conflict on two dimensions, namely circles with broad horizontal or narrow vertical stripes. The eye tracker records to which side the infant looks. (No reinforcing images appear in this phase.) Which side they look for each stimulus can indicate which dimension if any they have learned to use as the predictor.
This method offers the possibility of examining multiple dimensions of perceptual responding simultaneously and varying the test stimuli can examine how far the infant generalises. However a fair number of test trials are required to determine if the infant are reacting in consistent manner, exhibiting a bias or responding randomly. So test phase continues until each infant too fussy to continue, when it is possible they may unlearn the association. Then the analysis of results allows for different patterns of behaviour between infants. The analysis is complex and the results from it so far are unclear but the technique has potential to be a useful way of investigating infant categorization responses to complex stimuli.

1.15.3. Sequential Touching. Using highly simplified sorting tasks Riccuiti (1965) and Nelson (1973) both found evidence that when infants between 12 and 24 months were presented with a selection of eight objects they would sometimes group the objects but more reliably would sequentially touch items from the same class or category at a rate higher than expected by chance. Mandler and colleagues (Mandler and Fivush, 1987; Mandler et al., 1991; Mandler and McDonough, 1993) have developed this sequential touching paradigm to investigate global versus basic contrasts in early categorization and for testing distinction between perceptual and conceptual categories. Section 1.16.3 below discusses this work. This section describes the method itself.

An infant is seated at a table and presented with a selection of small manipulable toys, typically four from one category and four from another randomly arrayed on the table. An experimenter encourages the child to play with the toys, surreptitiously replacing any that fall to the floor. If child spends disproportionate amount of time playing with one item the experimenter encourages infant to play with others. Otherwise the experimenter does not engage with the infant. The experiment lasts several minutes. It may also involve a pre-test phase where infants are familiarized with some similar or related items to emphasize some aspect of the category structure being investigate. For example an experimenter may emphasize the movement or the parts of a toy to induce the infant to attend to motion or structure of the test items (Rakison and Poulín-Dubois, 2001). As the task requires a fair amount of manual dexterity it is only suitable for infants older than about 12 months.
The infant is filmed and the video later scored for the pattern of the infants’ interaction with the toys. One ‘touch’ being counted for each time a toy is manipulated with ‘active engagement and attention’ from the infant. But not if it is accidentally brushed or if infant is not looking or is banging the toy on the table without attending to its’ particular characteristics. This makes the scoring a subjective process but high inter-rater reliability scores suggest it is a valid measure. The dependent measure is the mean length of sequences of touches. If an infant is aware of the category groupings of the toys then they may touch items in a non-random manner. For example, playing selectively with all land mammals. The actual patterns of touches are compared to mean expected sequences generated by Monte-Carlo simulation of random touching.

Mean sequence length is not an ideal measure because it is not clear why infant would stay within a category or conversely that failure to do so indicates a lack of awareness of categories. Mareschal and Tan (2007) take a more sophisticated statistical approach to sequential touching data using finite mixture models (Thomas and Dahlin, 2000; Molenaar et al., 2001) to test for possibility that different infants may have different touching patterns. For example, some infants may avoid items not in category A whilst others may alternate between two categories A & B. Traditional analysis could not discriminate these strategies and could potentially miss instances of category learning.

One further interesting observation about the sequential touching methodology comes from one of the earliest investigations into its use. Starkey (1981) examined sequential touching in 6, 9 and 12-month old infants using a range of abstract stimuli, small brightly coloured objects that could differ in size, colour and shape, or in combinations of these properties. Infants were tested for grouping and sequential touching with sets of 8 objects consisting of 2 groups of 4 objects of varying discriminability. He found some sequential touching in 94% of 9 month olds and 100% of 12 month olds. Furthermore, he found that the greater the distinctiveness, that is, the more dimensions the groups differed on, the more likely the infants would be to demonstrate sequential touching. While this result might seem obvious, it suggests that sequential touching might be a suitable methodology for quantitative investigation of infants’ awareness and attention to multiple stimulus dimensions within categorization.
1.16. Categorization in Infancy - Experimental Findings and Developmental Theories

There are many experimental results concerning infants’ abilities to categorize stimuli during the first year of life. These have been well documented and categorized by researchers. Almost from birth infants possess a range of perceptual constancies, treating as perceptually equivalent objects that differ visually in size (Granrud, 1987) or in shape (Slater and Morrison, 1985). By four months old infants can form perceptual prototypes for abstract stimuli (Bomba and Siqueland, 1983) and perceptually driven natural categories of pictures of animals (Quinn et al., 1993) that are very sensitive to distributional information (French et al., 2004). At around six months they begin to show categorical responses to abstract spatial relations like containment and support (Casasola and Cohen, 2002) or ‘above’ and ‘below’ (Quinn, 2003). By about 10 or 11 months they are sensitive to feature correlations (Younger and Cohen, 1986) and capable of forming and utilizing abstract conceptual categories (Mandler and McDonough, 1993).

Despite or perhaps because of this wide range of results there has been less success in fitting the data into unified theoretical frameworks. Over and above these findings in the categorization literature there is a very broad domain of results that could be included under the rubric of categorization and concept formation: Almost all of infant cognition in the first year is about the acquisition and development of concepts and knowledge of objects and events in the physical world. As such questions about infant category learning is inextricably entangled with questions concerning the development of perception, object knowledge, causal knowledge and other early cognitive skills.

In contrast to the adult literature where category learning and category knowledge are well demarcated, the distinction is rarely explicitly made in the infant literature. This, in itself, is a theoretical problem but it also interacts with another difficulty for developmental theories, namely the seemingly contradictory and highly task dependent of patterns of categorization found in the first year. Where, for example, basic level distinctions are present at 3 months with looking time measures but not before 11 months using object handling paradigms. Largely, these are false
1.16. CATEGORIZATION IN INFANCY

Figure 1.9: Examples of prototypes and levels of distortion used in Younger and Gotlieb (1988). The rows represent different levels of 'goodness of form' or greater symmetry. The columns represent increasing amount of distortion, generated by randomly shifting each point increasing distances away from the prototypical form.

oppositions due to imprecise formulation and separation of terminology and problem spaces. Clearly and in contrast to adults infants are continually learning and updating their current knowledge. Likewise infants perceptions and representations will be very different at different ages and in different situations. But without a rigorous definitional framework this is often overlooked. In particular, by leaving out a clear account of the interaction of category learning and category knowledge in a given task or situation, any theoretical explanation will be necessarily limited. One advantage of the connectionist modelling approach to infant categorization is that a good model must explicitly address both questions; the processes whereby infants acquire categories and the structure of their encoded category knowledge. See section 1.16.4 below.

1.16.1. Categorization and prototypes. Several early studies were motivated by the findings of Posner and Keele (1968) that adults make use of perceptual prototypes. Using similar sets of distorted dot patterns researchers have investigated infants’ ability to extract perceptual prototypes (e.g. Strauss, 1979; Bomba
and Siqueland, 1983; Younger and Gotlieb, 1988.) Figure 1.9 illustrates some typical stimuli used. Infants are usually tested using the visual novelty preference measures described in section 1.15.1.1 above, which means they will only be shown a very limited number of exemplars, typically no more than a dozen and often only three or four exemplars show several times. Category structure must necessarily be less complex than for adult participants. Nonetheless, the studies tend to support the view that infants treat these simple visual perceptual stimuli in much the same way as adults do.

Bomba and Siqueland (1983) presented 3- to 4-month old infants with familiarization to multiple distortions of a single given prototype (square, triangle, etc.) at different levels of distortion as per figure 1.9. The unseen prototypes were paired with a previously seen distorted exemplar and infants visual preferences were recorded. When familiarization was with a small number of exemplars no preference was found. Increasing the number of exemplars or introducing a delay between familiarization and test cause the infant to look longer at the previously seen exemplar over the prototype. Indicating that, like adults, they had extracted and stored some kind of prototypical representation.

Quinn (1987) sought to investigate if infants can still extract prototypical categorical representations when distorted exemplars are presented with distracting non-categorical exemplars or with items from a contrasting category. An experimental situation more real life where infant must learn categories in the face of wide variation in the stimuli in the world. Moreover, he hypothesized that contrasting information may actually aid the formation of categories by emphasising the relevant features that discriminate one category from another. Testing 3- and 4-month old infants, Quinn found support for this hypothesis. Distracting non-categorical exemplars did not hinder category learning and infants were able to learn two categories simultaneously when presented with familiarization to exemplars of both at once.

Both the above studies demonstrated categorization in very young infants but only using very simple gestalt forms with high degree of symmetry. To test for developmental differences in infants’ categorization abilities Younger and Gotlieb (1988)
tested 3-, 5- and 7-month olds with dot patterns of increasing complexity, decreasing 'goodness of form' as judged by adult raters. Whilst 3-month old infants only showed evidence of categorization with the simpler forms, older infants extracted prototypical structure in the increasingly more complex patterns. While consistent with the previous studies this result is important as it emphasizes the need for systematic use of comparable stimuli to determine the changes that take place in categorization ability with development.

1.16.2. Features versus Correlation. When the stimuli and the possible ways of categorizing them are more complex different strategies may be revealed. In Strauss (1979) infants were reported to have extracted a median prototype from a set of 14 schematic faces that varied on 4 dimensions. Whereas in Sherman (1985) infants treated as more familiar a modal exemplar composed of the most common individual features from a set of 4 schematic faces varying in 3 dimensions. In the former case it seems that infants are using some abstracted prototype, whilst in the latter they may be identifying the face by encoding individual features separately. A series of studies by Barbara Younger and colleagues (Younger, 1985; Younger and Cohen, 1986) investigated developmental differences in infants’ ability to make use of overall feature similarity and/or correlations between feature occurrence.

Younger (1985) demonstrated that not only do infants detect violations of feature correlation but that they can, by 10-months-old, use correlations to partition a set of stimuli into several categories. Figure 1.10 illustrates how the individual features of an artificial stimuli can be be varied. Using figures of this type, two sets of stimuli were created. In one unconstrained set $A$ each feature could take either a large or a small value separately to other features. In the other set $B$, the values co-varied, half the animals had long necks, short legs, thick tails and closely paired ears, whilst the other half had short necks, long legs, narrow tails and widely spaced ears. Infants who had seen exemplars from set $A$ formed a single category. They did not dishabituate to an average stimulus with medium values on all features but did look longer at exemplars with similar patterns of features to animals from set $B$. Conversely, infants who had initially been exposed to set $B$ dishabituated only to the average stimulus, indicating they had formed two categories on the basis of the correlational information.
Younger and Cohen (1986) used a visual preference paradigm with line drawings of imaginary animals which could be shown having different combinations of bodies, legs, feet, tails and ears. Infants were shown exemplars that were constrained such that certain kind of body, tail and feet always went together. Four-month olds processed the features individually and independently, dishabituating to a test exemplar having novel features but not to one have previously seen features in novel combinations. Whereas 10-month olds dishabituated in both case, indicating a sensitivity to feature correlations. Like the 4-month olds, a group of 7-month olds treated correlated and uncorrelated equivalently. However, in a discrimination test limited to just two animals they were able to detect a correlation violation suggesting failure to categorize was possibly due to memory limitations. These intermediate results and analysis of modelling results (see section 1.16.4.2) indicates that the developmental progression from feature to correlational could be due to multiple factors. The development of memory, perceptual and representational
acuity all interacting to provide infant with ability to detect correlations (Younger et al., 2004).

There are limitations with the above studies. Firstly, as indicated there may be multiple developmental factors involved but these relatively simple paradigms do not provide a means of differentiating these. Secondly, the stimuli are relatively stylized and perceptually impoverished. Not only is there difficulty in determining if infants will pay sufficient attention to learn the correlations but also the measuring of similarity and difference between the 'average' and non-average stimuli is somewhat arbitrary, sometimes relying on adult ratings, sometimes just using the scalar dimensions of the features. Finally, no general conclusions can easily be drawn because there are many task and stimulus dependencies on the age at which infants categorize different objects in different contexts. For example, using conditioned leg kick, infants as young as three months have been able to integrate colour-form relations (Bhatt and Rovee-Collier, 1994) whereas using a paired preference looking task Younger (1993) found that 7 months olds but not 4 month olds could make colour-form integrations. Further discussion of this developmental progression from featural to correlational processing and task dependencies are to be found in Younger (2003).

1.16.3. Conceptual Primitives. It is widely accepted that by the second year of life infants can utilize abstract conceptual knowledge. They can group or discriminate global categories such as animals and vehicles even when the individual exemplars are themselves perceptually diverse. They have expectations about hidden properties of objects of a particular class, for example they know that animals eat, sleep and can move by themselves, but that items of cutlery do not. However, there is a major debate as to whether these abstract properties are deducible from perceptual inputs alone or if the so-called perceptual to conceptual shift requires meaning to be imposed with reference to conceptual primitives or 'image schemas'.

This latter position is argued by Mandler (1992; 2000; 2003). She believes that perceptual learning mechanisms alone cannot provide sufficient information to allow the infant to solve conceptual categorization problems. She presents data from object manipulation experiments that show that infants as young as 11 months can classify using abstract categories. Moreover, she claims that global categories
appear before basic categories. (Global categories are seen as precursors to adult superordinate categories, the different terminology is deliberately chosen to emphasize that the infants’ categories are likely to be quite different from their adult equivalents.) Mandler et al. (1991) tested 19 to 30-month-olds using sequential touching paradigm and Mandler and McDonough (1993) used an object examining task to extend testing down to 9 to 11-month-olds. In both studies, infants discriminated at a global level (e.g. animals vs vehicles) but only did so for some basic level pairings, usually when the contrast between them was high (e.g. cars vs. aeroplanes but not cars vs trucks.)

This leads to the criticism that infants are not solving the global tasks by conceptual means but are relying on some less obvious perceptual regularities or differences. So although birds and aeroplanes are similar in shape, it could be that birds get grouped with animals because both have eyes or similar surface textures and organic contours and whilst aeroplanes and vehicles have more flat surfaces and are made from similar materials. A study by Pauen (2002a) attempts to address this criticism. Using an object examining paradigm, groups of 10 and 11 month old infants were familiarized with either model animals or furniture. In the critical conditions the models were carefully designed to have very similar appearance across the categories; all models were made from wood, with similar shapes and colourings with doorknobs serving as eye-like markings on the furniture. The between-category similarity did not affect categorization. However, here the conclusion that categorization was based on prior knowledge relies on a null finding. This is an interpretive problem that also affects conclusions derived from infants’ failure to categorize by basic properties in certain tasks but not others.

Alternative accounts dispense with the conceptual primitives and claim that conceptual categories are the product of a process of perceptual enrichment which may or may not be driven by the initial salience of certain perceptual properties. Quinn and others (Quinn and Eimas, 1997; Quinn and Johnson, 1997; Mareschal, 2000) believe that early perceptual experience provides the basis for later more conceptual or knowledge driven representations. Knowledge of hidden or rarely occurring properties can be acquired by associative learning mechanisms. So that, for example, photographs, toys and actual dogs or cats all share many common features (eyes, ears, fur, legs, etc.) that go to defining or building a common category structure.
Other information that is not always shared or experienced, such as how the animal moves, that it barks, etc. is associated with this structure and can also be recalled because of this association. Likewise, properties that are shared by broad group of objects or events will tend to be learnt before specific details, giving a process by which abstract concepts are built. For example, in the simulation in Quinn and Johnson (1997) a global level emerged before basic level, consistent with findings of Mandler and McDonough (1993).

Rakison’s proposal (Rakison, 2003) is slightly different, in that it emphasises the initial salience of some perceptual features and hypothesizes that it is the infants’ attention to these that drives the enrichment of infant categories. In contrast to Mandler’s idea of abstract image schemas for different types of movement, Rakison observed that infants are likely first notice that some objects have wheels and others have legs and yet others have neither. Infants’ attention would be captured by these highly salient features, they would observe that objects with legs and those with wheels moved in different ways and this global distinction would be the basis for associating other features with these categories, from whereon the enrichment proceeds as per Quinn’s account. Using sequential touching and object manipulation measures, Rakison and Butterworth (1998) found 14- to 22- month olds attended to and formed categories by parts, treating objects with legs differently from those with wheels, even when these objects were artificial half animal, half vehicle hybrids. Chapter 4 of Rogers and McClelland (2004) presents a connectionist account that could potentially reconcile these two positions by allowing that the initially salient features are precisely those that most commonly co-occur in early experience. See section 1.16.4.3 below.

1.16.4. Computational Models of Infant Category Learning. Several computational models of aspects of infant category learning have been presented. This section briefly reviews some of these. Here, we are less interested in how the specific details of particular models fit certain data-sets and more concerned with the general principles of how modelling can inform developmental research on categorization.

1.16.4.1. Asymmetric exclusivity. Mareschal et al. (2000) demonstrates how differing distributions of features across dimensions can account for asymmetrical
inclusions of one category within another; 3.5 month old infants and auto-associator networks exposed to broadly varying category of Dogs will generalise to include novel cats, but the less varied category of Cats excluded novel dogs. The model was able to successfully predict that memory and interference effects would be observed in variations of the task. While a subsequent study (French et al., 2004) was able to reverse the direction of the asymmetry in infants by manipulating the variability of the stimuli used according the predictions of the network model. In this case the sets of specific pictures cats and dogs were selected so that the Cats category was more diverse as measured by variation the perceptual parameters that the model operated on, the distributions on these dimension now formed an inclusion relation for the distributions of the same dimensions in the Dogs category.

This work is highly suggestive of the fact that infants are sensitive to multiple low level dimensions in perceptual inputs and that analysis of infants categorization performance in these 'bottom-up' terms can be revealing. It provides support to the perceptual enrichment account of the perceptual to conceptual shift by demonstrating that the information to form semantic discriminations exists at a perceptual level and that infants are paying attention to this information in some online category learning tasks.

1.16.4.2. Feature Correlations. In a special issue of Infancy in 2004, three different models were presented (Gureckis and Love; Shultz and Cohen; Westermann and Mareschal) that each attempted to explain the findings of Younger and Cohen (1986) discussed above in section 1.16.2. Following three very different approaches they all modelled the development changes in the ability to process correlations among features between 4 and 10 months. Connectionist models are associative learning mechanisms that are very good at detecting correlations so the challenge for these models was to fit to the very poor performance of young infants. The constraints placed upon the models having behavioural analogues that can be compared to the infant data and be used to generate further predictions.

Gureckis and Love used a model of adult categorization in one case manipulated a parameter equivalent to a restricted memory capacity and in another added noise to the input as an analogue to the reduced perceptual acuity of younger infants. Shultz and Cohen use a 'cascade correlation' network that can recruit extra resources to
fit the training data. By requiring model of 10 month old performance to fit the
data more accurately, it needed to use extra hidden units and was as a consequence
able to detect correlations. Shultz and Cohen argue that this is equivalent to
the greater processing capacity of older infants. Westermann and Mareschal use
an auto-encoder network with a Gaussian activation function on its' hidden units,
varying the width of the bell curve effectively changes the level of detail the network
could represent across the feature space with greater accuracy in representation
allowing correlations to be detected. Westermann and Mareschal put forward the
'Representational Acuity Hypothesis' stating that an equivalent tuning in feature
representation occurs in development.

In all likelihood all four mechanisms (memory limitations, perceptual acuity, repre-
sentational acuity and depth of processing) have real developmental equivalents
in early infancy and will have an influence on the developmental trajectories of
cognitive abilities.

1.16.4.3. Category Structure. In their book *Semantic Cognition* Rogers and
McClelland (2004) provided a detailed and in depth investigation of how a feed-
forward network with distributed conceptual representations can fit both develop-
mental and adult behavioural data on knowledge of taxonomic hierarchies. Albeit,
these are demonstrated with a toy model that learns very simplified sets of concepts
about natural world of the type:

{sunflower, salmon, maple, etc.} >

{HAS, CAN, IS, ISA} >

{legs, grow, yellow, animal, dog, etc.}

Nonetheless, the networks succeeds at modelling a wide range of experimental data.
From the emergence of global before basic concepts in pre-linguistic infants, through
typicality and basic level reaction time biases in adults through to the patterns of
knowledge breakdown found in semantic dementia. This breadth of coverage by a
single model is impressive in itself and the continuity between developmental and
adult accounts with actual qualitative data fitting and predictive power makes a
strong case for connectionist approaches to development.
Chapter 4 of Rogers and McClelland (2004) deals with concept acquisition in infancy. It addresses the debate about whether conceptual knowledge can arise directly from perceptual inputs or if it requires either representational primitives and schemas or innate knowledge of non-observable core causal properties. Rogers & McClelland suggest that their model goes beyond the limitation of previous perceptual associative learning schemes because its sensitivity to higher order groupings of correlations allows it to pick out similarities between perceptually varied stimuli. The model also provides a mechanism for inductive generalizations, whereby the coherent covariation of object features and properties within a category lead to the expectation that other related or similar members of the category will have similar properties. For example, because the representations of dogs and cats are very similar learning a new fact about dogs also effects the expectations that this fact will apply to cats. This is significant because Mandler cites her work inductive generalizations in imitation tasks (Mandler & McDonough, 2000) as strong support for the necessity of conceptual primitives. This model undermines that claim. While all these results are suggestive, this is only a toy model. Knowing whether such a mechanism operates in infancy requires further research into the range and saliences of infants perceptual inputs.

1.16.5. Categorisation in Infancy - Summary. As can be seen from the foregoing survey, there has been a great deal of research into category learning in infancy. It has found both prototype (Bomba and Siqueland, 1983) and exemplar effects (French et al., 2001). Furthermore, the studies of Barbara Younger and colleagues have established that young infants can detect feature change but not correlation change while older infants detect both. Researchers have not yet been directly concerned with rules, but they have been interested in the degree of abstraction of that is possible in representations. This question has been approached via the debate about the primacy of either global versus basic category representation. The debate divides along nativist/constructivist lines with Quinn and colleagues supporting a bottom-up driven by similarity and associations, while Mandler advocating a process driven by conceptual primitives. Although the existing literature on category learning in infancy currently has little to contribute to the rules versus similarity debate, it does over a wide variety of sensitive experimental techniques
1.17. Rules and Similarity in early childhood and beyond

1.17.1. Conceptual and linguistic development. If we look beyond the first two years of life then there is a sudden explosion of the concepts and topics that children appear to know about and which researchers will ask them questions about. Research reveals knowledge of superordinate or subordinate category structure (e.g. Mervis, 1994), the 'fast mapping' of new words onto objects (Carey, 1978) or onto categories (Waxman and Booth, 2000). It is no coincidence that this coincides with an explosion in children’s vocabulary and linguistic sophistication (McMurray, 2007). In part no doubt the recorded spurt in conceptual sophistication at this age is an artifact of these improved communication abilities which permits children to understand task instructions and allows researchers to look at usage patterns to determine conceptual awareness or ask children direct questions about the conceptual knowledge and which moreover the children can answer. But various theories explicitly and intimately link the development of linguistic and conceptual knowledge (e.g. Bloom and Keil, 2001; Waxman, 2003 or the volume edited by Bowerman and Levinson, 2001). For example, Waxman (2003) proposes that "words are an invitation to form categories" (p.220, Waxman, 2003) and the initial link between word and categories "sets the stage for more precise expectations linking particular grammatical forms (e.g. nouns, adjectives, verbs) to particular types of relations among objects (e.g. object taxonomies, object properties, actions)" (p.220, ibid.)

However, while these proposals are interesting, from the perspective of the current research, any link between language and category learning adds too many extra degrees of complexity to the problem under consideration. As was indicated in Section 1.16, our two primary motivations are comparisons of the respective rule- and similarity-based category learning of children and adults. Whilst, linguistic considerations may also effect the nature and structure of adult categories, the prospect of having to also account for variation in linguistic ability, dramatically reduces the feasibility of such a study. Therefore, for the remainder of this section and paradigms with which such questions could be approached. We return to the question of what might be investigated in section 1.18.
we shall restrict our interest to non-linguistically based tasks. In this respect, relational categories are of interest.

1.17.2. Relational Categories. Gentner and Medina (1998) investigated the relationship between rules and similarity with respect to relational categories. They give as an example, a study by Kotovsky and Gentner (1996) which investigated the abilities of 4, 6 and 8 year old children to recognize relational similarities. Children were presented with triads of objects that demonstrated some relation and were asked to pick the equivalent relation from two alternatives. The test items varied on dimension and on polarity. For example, if "xXx" were the initial configuration, the choices might be "oOo" or "ooO" in the unchanged case, "OoO" or "oOO" in the changed polarity case and "121" or "112" in the changed dimension case. Six and 8-year-olds were responding above chance in all conditions but 4 year olds were only capable of answering correctly when dimension and polarity were unchanged. However, providing them with coaching that drew attention to these dimensions without explicitly highlighting the commonalities improved their performance on the generalization task.

Gentner and Medina (1998) claimed that "the process of structural comparison can act as a bridge by which similarity-based processes can give rise to abstract rules." (p. 266, Gentner and Medina, 1998, emphasis added). Here they use the definition of rules from Smith et al. (1992) (see Box 1.5.5). Their theoretical model is described by a structural mapping algorithm which attempts to align one structure with another. Here a structure is a set of objects, together with their attributes, features, relations and higher order relations. Initially the algorithm detects any matches between any elements of the structures but by noticing consistent matches (or having them pointed out) the algorithm moves towards mappings that the elements of a structure in one-to-one correspondences. Hence creating and preserving the relations at an abstract level.

However, as an alternative to this explicit model, Leech et al. (2008) put forward a developmental connectionist model that succeeds on analogy type task by relational priming. Leech et al. argue that their model has several advantages over structural mapping in that it is developmental rather than derived from explicit adult theory and accounting for a wider range of developmental phenomena. Such
as the finding of Goswami and Brown (1990) that children’s analogical reasoning is
domain-specific rather than domain-general. This counts against the deep abstract
nature of relations claimed by Gentner and Medina (1998). Whichever model is
favoured, all these results suggest that there are can be interaction of abstraction
and similarity over development, making relational categories a potentially fruitful
avenue to explore.

1.17.3. From Associations to Rules? It has been suggested that over the
course of development there is a shift from associative, similarity based classifications
to classifications based on rules or single dimensions; namely that in spontaneous classification tasks young children would tend to sort items on the basis
of overall similarity across a range of dimensions whilst adults will often make a
binary discrimination on the basis of a single dimension (Smith and Kemler, 1977;
Smith, 1981; Smith and Kemler Nelson, 1984; Medin et al., 1987). Recall from
section 1.3.6 that in free sorting tasks adults are found to classify objects on the
basis of a single dimension (Medin et al., 1987).

Children are more typically tested with a triad classification task. Figure 1.11
illustrates some typical arrangements. Three stimuli A, B, C are presented simultan
eously, such that A and B are identical on the value of a particular dimension
whilst B and C are similar but non-identical to each other across a range of dimen
sions. Participants are asked which two items go together with the former pairing
giving evidence of dimensional responding and the latter indicative of similarity.
The inconsistent pairing A and C is sometimes disallowed and sometimes included
as a measure of guessing (see Raijmakers et al., 2004). Using this procedure with
squares of different size and brightness, Smith and Kemler (1977) found that chil
dren aged 4 to 7 would primarily group objects by overall similarity whilst older
children would rely on identity on a single dimension to group items. For exam
ple, young children would group a 2.54cm white square with a 1.27cm grey square
rather than with a 2.54cm black square, whereas older children would often group
the white and the black squares together because of their equal size. Smith (1989b;
1989a) interprets this as a consequence of the development of awareness of "kinds
of similarity", namely the ability to be aware of identity or other relation (greater
1.17. RULES AND SIMILARITY IN EARLY CHILDHOOD AND BEYOND 113

Figure 1.11: A range of possible triad types from Thompson (1994), experiment 1. Participants are shown stimuli A, B and C which vary on two dimensions as indicated and asked which two go together. In triads of types Ia, Ib, IIa & IIb a A-B pairing is taken as a dimensional classification, B-C as a similarity classification and A-C as an inconsistent outcome. Whilst in types IIIa, IIIb and IV the groups are classified by whichever dimension is chosen. (Diagram taken from Raijmakers et al., 2004)

than) on a single dimension. She even suggests that this "may be one of the major intellectual achievements of early childhood" (p.146, Smith, 1989a).

However, other developmental researchers are less clear that there is in fact a holistic-to-analytic shift. Ward et al. (1990) conducted a set of experiments with groups of 5, 7 and 9 year old school children and college age controls using stimuli that could be classified by any of four dimensions or by an overall family resemblance structure. They found that all age groups primarily used analytic classifications based on just one or two stimulus dimensions. The proportion of accurate analytic classifiers was much lower in the younger age groups but there was no evidence of any holistic classifiers at any age. This result is consistent with the differential-sensitivity theory of Cook and Odom (1992). They propose that participants of all ages can perceive the different dimensions and will preferentially use dimensional rules but that younger participants will do so inconsistently and are less sensitive to dimensional differences and that relative saliences in the dimensions will change with development.

Thompson (1994) used a rule analysis methodology to assess the performance of 4-to 10-year-olds and college age controls on a triad classification task with discs that
varied in size and brightness. She found support for the differential-sensitivity theory with participants consistency increasing with age. However, Raijmakers et al. (2004) criticise the arbitrary assumptions of Thompson’s rule analysis technique (the need to a priori specify the number and type of possible rule together with arbitrary decision level for deciding if a rule is being used). Instead they introduce latent class analysis as a statistically more rigorous approach. Rerunning a size-brightness task similar to Thompson (1994) and a size-orientation task with 4- to 12-year-olds Raijmakers et al. (2004) also find that the differential-sensitivity hypothesis is supported, with no evidence of wholistic classification at any age. However, these experiments all used very simple stimuli with only a small number of dimensions that often took only a small number of alternative values, thus it might be argued they are too simple or have certain demand characteristics that make holistic responding unlikely.

1.17.4. Summary. This section reviewed literature relating to the differences between rules and similarity found in early childhood. Initially, we saw that the development of concepts is likely to be closely tied to the development of language. However, this research was not considered in detail because of additional difficulties that linguistic development would bring into any comparison of children’s and adult’s category learning. The implication being that the development of rule- and similarity-based categories ought to be investigated with non-linguistic stimuli. Relational categories were then considered as an example of the use of very simple rules where similarity and abstraction appear to interact. Finally, a range of studies using the triad-classification task were considered. The evidence does not yet decide between two competing developmental accounts, the holistic-to-analytic shift theory and a differential sensitivity theory. Primarily because of the limitations of the triad task itself.

1.18. Conclusions and questions for Research

As was stated at the outset, this research intends to investigate the early development of Rules and Similarity by means of three approaches. By looking for the earliest available evidence of rule use, by comparing infants’ responses to rules and similarity directly in a single task and by making direct comparisons of children
and adults on rule and similarity based tasks. Prior to specifying how those three objectives would be met, an extensive literature review was undertaken. We started by surveying two recent theoretical accounts of the rules and similarity distinction from the perspective of the adult literature. We then examined the nature of rules from many different perspectives within cognitive science as to better guide our intuitions about what rules are and how they might develop. A shorter time was spent considering some of the difficulties with similarity and looking at the mechanics of the Generalized Context Model. Before a final review of some of the methods available and existing findings in previous developmental work on category learning. In this final section, we shall try to organise the questions raised in the course of the review according to the three objectives set at the beginning. In the hope that this provides clear motivation and direction to these themes.

1.18.1. Infant pre-cursors to rule. Two very interesting and as yet unanswered questions are; What is the earliest that rule use can be found and what form will the earliest rules take? Tied up with this question are considerations about what rules would available infant testing methodologies be capable of detecting. Ashby and O’Brien’s (2005) neuropsychological theory appears to predict that infants cannot use rules. By Pothos’s account infant rules need not be any different in form from adult rules. But how would an infant’s rule measure up to the criteria of Smith et al. (1992)? How would the criteria that learning is abrupt or that participants expect to show awareness of their rules (Reese, 1989) apply to infants? Or how might we assess the certainty of a rule or any abruptness of learning given the inherent noisiness of infant data? Not many techniques of infant testing produce learning data, should we try an unproven paradigm like the tracking of anticipatory eye movements (McMurray and Aslin, 2004)? Do any other insights help at this early stage? Is simplicity (Chater, 1996) appealing to infants? Perhaps the animal literature will enlighten us?

To help make this more concrete we can observe that the current earliest available evidence for rule use appears to come from study of artificial grammar learning (AGL) tasks (Marcus et al., 1999). Therefore we propose examining pre-cursors to explicit rule use in pre-linguistic infants by using artificial grammar learning
paradigms, paying particular attention to the learning process itself and investigating whether statistical mechanisms give rise to rule-like performance. This is the aim of Chapter 2.

While the relational categories of Gentner and Medina (1998) might be some of the simplest rules. Furthermore, Gentner and Medina (1998) claimed that rules and similarity can be related the process of comparison and use this to motivate their investigations of relational categories so this may be a productive avenue to explore. Therefore we are interested in answering the question, what is the simplest relational property that we demonstrate abstract learning in infants? This is the topic of Chapter 3.

1.18.2. **Comparing rules and similarity in infants.** The second question asked if there is any competition between rules and similarity in infancy? Is the PFC likely to involved at this young age? Is infant categorization flexible? The infant category learning literature leads us to expect that infants detect single feature change and correlational category structure (Younger and Cohen, 1986) Would this map onto a rule and similarity distinction? Or would top down and bottom up effects be manifest as rules or similarity? Again there are questions about what are the appropriate techniques and choices of stimulus material. And by what criteria should learning and discrimination be assessed? Chapter 4 investigates these questions.

1.18.3. **Comparing children and adults.** Our final area of interest concerned the continuities that may exist between children and adults. What are the continuities and what are the differences? Are these the consequence of a single process that becomes more reliable with age. Or do developments in the PFC allow for more complex representations to be manipulated and learned and treated as rules?

Thus, children and adults will be tested in identical fashion on supervised and unsupervised category learning tasks using categories with either rule-based or similarity-based structures. The intention is that the tasks will be simple enough for all participants to understand but sufficiently challenging that any variations in abilities may be found, either between different ages, between different category
structures or as interactions between these two variables. The tasks will use unfamiliar artificial materials that control for prior experience and minimise the verbal component of the task. They will be presented in the context of a novel specifically designed computer game that aims to present a task with a high degree of ecological validity. These issues are discussed in greater detail in Chapter 5.
CHAPTER 2

The role of stimulus complexity in visual sequence learning by five month old infants and adult controls.

2.1. Introduction

In this chapter we describe several experiments conducted to investigate the response of adults and young infants to visually presented sequences of information and attempt to assess what aspects of the stimuli they are learning or otherwise responding to. We begin by surveying the reasons why researchers have used artificial grammar learning (AGL) paradigms with infants. In particular how AGL are thought to provide a means of investigating early aspects of language acquisition and other rule use. We summarise some of the main empirical findings and theoretical perspectives on these results. Additionally, we introduce the idea that AGL paradigms can also be used to investigate infants’ awareness of and response to the relative complexity of stimuli they encounter. We briefly survey several mathematical formalisms by which the complexity of artificial grammars could be assessed and use these to analyse the data from our experiments. All the experiments build on visual sequence learning paradigm introduced by Kirkham et al. (2002). Three experiments with five-month-olds investigate their relative interest in sequences of looming coloured shapes with varying statistical properties whilst a further experiment adapts the method to provide a control task suitable for adults.

2.1.1. Artificial grammar learning in infants. For many researchers language is the quintessential example of rule use in humans. This view is most clearly associated with the continually evolving position of Chomsky (1957; 1965; 1993). The rules in language are subdivided into independent rules of syntax and morphology and to a lesser extent the rules of phonology. In Words and Rules (2000), Pinker identifies the productive, combinatoric and symbolic aspects of language as
evidence of its’ rule like nature. From a different perspective, within the categorization literature Ashby et al. (1998) equate being rule-like with being verbalizable. They suggest that rule-like categories are precisely those for which it is possible to give a simple verbal description of the criteria for category membership. See section 1.6 for a wider survey.

One matter of debate in the study of language is whether the detection of patterns and regularities is due to a specific human adaptation for language or a more general set of pattern recognition mechanisms in the brain. Clearly it is not possible to resolve this question by studying adult native language performance. Developmental approaches and controlled laboratory experiments with novel materials are required. Since their introduction by Reber (1967), artificial grammars have provided a means for investigating the mechanisms of language learning in relative isolation from the uncontrollable influence of prior experience. Participants are exposed to a wholly unfamiliar and highly constrained ‘language’ and assessed on their learning of certain critical features of that language. The artificial language can take the form of strings of letters or artificially synthesised syllables, tones or visual cues. Participants can be assessed on their recognition of previously encountered items or generalisation to novel items following the same grammatical rules, assessed by either changes in reaction times with learning or on their accuracy at making grammaticality judgments.

Numerous studies have claimed to show that infants can learn regularities in artificial grammars (Jusczyk and Aslin, 1995; Safran et al., 1996; Gómez and Gerken, 1999; Marcus et al., 1999). In particular, infancy researchers have used artificial grammars to investigate the early precursors of the combinatoric and symbolic aspects of language. Jusczyk and Aslin (1995) demonstrated that infants can detect word boundaries in fluent, continuous speech. Safran et al. (1996); Aslin et al. (1998); Safran et al. (1999); Safran (2001) tested the hypothesis that humans possess a learning mechanism that can detect these statistical patterns in the absence of other cues. Safran et al. (1996) presented 8 month old infants with monotone synthesised stream of syllables as a continuous recording lasting several minutes, like so

...tupirobidakutupiropadotibidakugolabutupiropadotigolabubidaku...
In fact, this was a chain made from four nonsense words \{tupiro, golabu, bidaku, padoti\} selected randomly (but with no word repetition). Infants were tested to see if they had extracted these statistical co-occurrences by comparing their reaction to 'words' \{tupiro, golabu\}, non-words \{dapiku, tilado\} or part-words \{piropa, kugola\}. With a conditioned head turn procedure, infants of eight months old were able to distinguish between 'words' they had heard before and 'non-words' consisting of novel combinations of the same syllables. They were even able to discriminate the part words, indicating that they had in effect created word boundaries in the continuous stream. These type of studies provide evidence that infants can quickly determine certain combinatoric aspects of language-like stimuli from statistical regularities alone (the synthesised speech had no stress, intonation or rhythmical cues).

The discriminations in these experiments were purely statistical in nature and so they do not address infants abilities to abstract relational features from an artificial grammar nor to generalise across inputs.

However several experiments have shown that infants can learn more than just simple transitional probabilities. In Marcus et al. (1999), 7.5 month old infants familiarized with AAB patterns preferred ABB patterns when tested with novel materials. In Gómez and Gerken (1999), Experiment 4 one year old infants exposed to one grammatical structure preferred this grammar over a novel grammar at test. There were no syllables in common between learning and test so infants could only be discriminating on the basis of grammatical structure. In Gómez (2002) 18-month-olds could learn multiple non-adjacent dependencies when there was a sufficiently high degree of variability in the intervening elements. Thus they could learn that \( pel-X_i-jic \) and \( vot-X_i-rud \) were both grammatical independent of the intervening two syllable words \( X_i \) when these varied over 24 possible items but not when there were only 2 or 12 possible \( X_i \). All of these findings have been replicated and extended (for example, Newport and Aslin, 2004; Gomez and Maye, 2005; Marcus et al., 2007; Saffran et al., 2007). The ability of infants to do more than just extract co-occurrence statistics is not in question but there is still debate as to how these results should be interpreted. In the next section we address some theoretical interpretations of AGL results and discussions of how they should be related to questions of language acquisition or debates about rules, similarity and associative learning.
2.1.2. Theoretical perspectives on Artificial Grammar Learning. In a review article Gómez and Gerken (2000) focus on AGL and its relation to infant language acquisition. They were interested in whether such research can address questions about the need for innate language acquisition devices versus the domain general learning and, additionally, how AGL tasks can address different aspects of the language learning problem. They identify four areas in which AGL has been applied as a research tool to the problems faced by an infant language learner. These are the problem of segmenting words from a continuous input stream. The necessity of encoding and recalling word order. The generalisation of grammar-like relations beyond the surface structure of the input, which they refer to as pattern-based abstraction. Lastly, the ability to learn that constraints on word orderings come about because of the words’ membership of syntactic categories (e.g. noun, verb, etc.), which they term category-based abstraction. Gómez and Gerken believe that results relating to the first two questions can be explained in purely statistical terms. Especially when one considers that infants’ natural learning environment provides a rich range of statistical cues in its prosody and phonology (e.g. Kelly, 1992; Morgan et al., 1996).

Gómez and Gerken believe that the latter two processes of pattern-based and category-based abstraction are the most crucial to the acquisition of a fully fledged natural language. They believe that Gómez and Gerken (1999) and Marcus et al. (1999) are both examples of perceptually bound pattern abstraction and that language learners can achieve this level of performance by making use of similarity processes and physically instantiated relations between elements, such an awareness of element identity when elements are repeated. They contrast this with the position of Marcus (1998a; 1998b) who argues that associative mechanisms could not in principle solve these problems. Although Altmann (2002) provided a demonstration of an associative learning neural network which could learn the Marcus et al. task this debate is far from over, as we discuss below. Additionally, Gómez and Gerken argue that there is a crucial distinction between pattern-based and category-based abstractions in that only the latter involve operations over genuinely abstract variables. There is no similarity metric that would identify a particular word as a noun or a verb but one needs access to this level of description to be linguistically productive. Knowledge is required at higher level of rules of grammar that can...
be applied to words of the appropriate class in order to correctly use novel words. Such that when learning a new noun _wug_ and a new verb _dax_ allows one to combine them in novel meaningful phrases such as "Three _wugs daxed_" and avoid grammatical errors like "Daxes wugged". Gómez and Gerken observe that children rarely make productive errors that cross category boundary such as pairing a determiner (the, a, an) with a verb. However, these high level linguistic considerations are beyond the interests of the current chapter where we are focusing on the lower level statistical side of infant artificial grammar learning. We therefore move onto other work that considers what is being learnt in statistical learning paradigms.

In a review article, Perruchet and Pacton (2006) highlight the contrast in emphasis between researchers using AGL to investigate implicit learning in adults (see Shanks, 2005 for a recent review) and those investigating statistical learning in infants, such as those described above. Perruchet and Pacton (2006) suggest that the former have focused on the acquisition and formation of ‘chunk’ knowledge while the latter have mostly focused on the extraction of transitional probabilities. They go on to make the claim that in both cases the mechanisms underlying performance are the same domain general incidental learning processes. Perruchet and Pacton start by downplaying the rules versus non-rules debate and do not question the findings of Marcus et al. (1999) but suggests the transfer could be due to analogical processes and take Conway and Christiansen’s (2002) finding of differential adult performance across auditory, visual and tactile modalities on Gómez and Gerken (1999) style tasks as evidence that the grammars are not processed at an abstract level. They go on to consider how chunk-based and statistical approaches to AGL could be combined and note that proponents of the statistical learning perspective have criticised the chunks approach because it appears to be equivalent to learning co-occurrence statistics. At this point, they emphasize the crucial finding that infants are learning transitional probabilities rather than co-occurrence frequencies (Aslin et al., 1998). For example, it is not so important that A and B occur frequently together but more useful to notice that B always follows A. However, Perruchet and Pacton counter that chunking models can capture this distinction if memory interference effects are built in, such as in Perruchet’s own model, _Parser_ (Perruchet and Peereman, 2004). They argue that when this approach is taken the differences between the two positions are much reduced and that future research
should focus on trying to account for learning of non-adjacent dependencies and how consciousness and attentional processes interact with incidental learning.

Pothos (2007) carries out a more extensive survey of theoretical positions on AGL, which makes it clear that there is a greater range and complexity to the views of AGL than are addressed by either Gómez and Gerken (2000) or Perruchet and Pacton (2006). Although he primarily focuses on the adult AGL literature the theoretical issues are equally relevant to infant tasks. He approaches AGL in terms of how it relates to and can inform two major theoretical questions in cognitive science, identifying two orthogonal dimensions along which AGL findings may be assessed. The first relates to the form of the knowledge the learner possesses, making the contrast between rules on the one hand and similarity or associations on the other. The second relates to how knowledge is processed during the task, making the contrast between implicit and explicit learning and cognition. In both cases the difficulties come down in part to the problems of giving meaningful definitions to the notions of rules and implicit knowledge respectively and in how these terms can be applied and methodologically assessed in the context of AGL tasks. Pothos considers these problems while surveying nine theoretical approaches to artificial grammar learning: rules, microrules, fragments, chunks, specific similarity, generalized context model similarity, simple recurrent networks, PARSER, and fluency. We refer the reader to Pothos’s review for specific details and references for each of these theories, here we shall just give a summary of Pothos’s assessment of the more general question of AGL taken as a whole.

This can be justified on the basis of Pothos’s conclusions that no single theory of AGL addresses all the data or satisfactorily answers the questions about the form of knowledge in AGL tasks or if that knowledge is implicit or explicit. Although he believes a synthesis built from existing theories may be possible if these core issues are addressed. To give clarity to his analysis of the short-comings of the existing theories he defines criteria by which the two dimensions may be assessed. To discriminate the implicit-explicit distinction he adopts the following definition which he attributes to Dienes (2004) "Implicit knowledge is knowledge not consciously activated at the time of a cognitive process" (p. 230, Pothos, 2007). While he adopts the frequency-independent conception of rule knowledge previously developed in Pothos (2005). From here he criticises the 'rule' theories of AGL for
making the assumption that participants’ judgements of grammaticality are equivalent to possessing psychological rules for the stimuli they have encountered. A microrules approach is better in that the rules are given by the participants and resemble the critical feature definition of rules from the categorization literature. The fragments approach is at some level equivalent to a set of microrules about bi-grams and trigrams. In contrast the remaining theories (with the possible exception of Parser) are all associative theories that differ primarily in details of their implementation. This seems to support data that similarity influences grammaticality judgments (e.g. Vokey and Brooks, 1992). A further question arises as to whether these mechanisms start from small fragments and work up or start from whole strings and work down but Pothos shows that in all cases these models arrive at frequency dependent expectations of what are the grammatical [n]-grams. Thus he concludes that it is frequency-independence that is the critical distinction between rule and similarity approaches to AGL and that future research should focus on this distinction. There is less clarity or consensus on how the theories could address the question of implicit knowledge, with a range of possibilities dependent on which cognitive mechanisms are thought to be involved at the point of learning or recall. The main conclusion here is that judgments that require large fragments would be beyond the capacity of working memory and would therefore necessarily involve implicit knowledge.

From this brief survey of theoretical work on AGL, we can see that there are several interesting unresolved questions and suggestive directions for research. The issue of domain-specificity is still very open. The successes of associative models of AGL suggest that domain general mechanisms may account for adult performance on AGL tasks. But it is less clear if these models are directly applicable to AGL in an infant language learning context because even though the infant tasks are generally simpler than those in the adult literature there is no necessity that the same mechanisms operate in both cases. Related to this is the question of prior experience and the extent to which it can influence performance either by shaping expectations about what patterns may be present or altering the salience of particular aspects of grammar. For example, analogously to the differences between adults and children on categorization tasks, where adults are said to favour low-dimensional rules
whilst children favour overall similarity, it may be that similar distinctions can be found between infant and adult performance on artificial grammar tasks.

A more general point, alluded to in passing by both Gómez and Gerken (2000) or Perruchet and Pacton (2006) is that great care is required in determining when infants exhibit analogous behaviours to adults. Testing infants requires simplified experimental designs which may alter task demands so far as to allow use of alternate mechanisms or even change the nature of the learning problem under consideration. For example, there is no simple way to consider the question of explicit or implicit processing with infant subjects. The question may even be meaningless in this context. This point may allow us to sidestep the considerable range of issues that Pothos’s extensive review of adult AGL literature has raised concerning implicit and explicit knowledge. Nonetheless, we should not neglect the fact that infants are generally actively engaged with the stimulus materials in infant AGL tasks and that considerations about their smaller working memory capacities should be kept in mind. There are also various unanswered questions about what is being generalized in transfer tasks where familiar grammatical structure is recognized with novel materials. We may accept Gómez and Gerken’s distinction between pattern-based and category-based abstraction and wonder about what types of relations between elements are salient or perceptible to an infant. This suggests that stimuli should not be treated in the wholly abstract sense of pure symbolic token (e.g. A, B, C, etc.) but a consideration should be made of all the perceptual qualities (e.g. loudness, tone, etc.) that an infant learner might have access to. Furthermore, we see from Pothos’s review that many models have different views of what counts as similarity between [n]-grams and how it may be used to learn or discriminate the grammar. Finally, given that frequency-independence is a crucial discriminant between rule-based and associative mechanisms in AGL, we should be careful to take account of the raw frequencies of items in the data-set and should look for expect comparable levels of performance with completely novel materials as evidence of rule use if we are looking at tasks featuring generalization and transfer.

2.1.3. Kirkham, Slemmer and Johnson (2002) revisited. With these questions in mind, Kirkham, Slemmer and Johnson (2002) is an interesting experiment to consider. The experiment familiarized infants with a very simple grammar.
using the visual presentation of coloured shapes and found that infants at 2, 5 and 8-months-old preferred a novel random sequence at test. The authors suggest that this result provides evidence for the domain generality of statistical learning in infants. This experiment also used one of the youngest age groups to have been tested with AGL paradigms. However, it also has several shortcomings and leaves a range of alternative possible mechanisms unexplored. Here we briefly describe that experiment and its results and go on to consider how a simpler explanation in terms of stimulus complexity could account for the pattern of findings. In turn, this will lead us into a consideration of how complexity of a given grammar may be measured.

The Kirkham et al. (2002) study was a simplified replication of Saffran et al. (1996) in the visual domain. The stimuli were looming monochrome coloured shapes that appeared one after another on a black background. Infants were habituated to a structured grammar and then tested for looking to this versus a novel random sequence. On each trial of the habituation phase infants were exposed to a continual stream of colour shapes each of which grew in size in the centre of the screen before being replaced by the next. The habituation phase ended when there was 50% decrease in looking averaged over a four trial window or when the infant had seen 12 habituation trials. Each trial ended when the infant looked away for 2 seconds or when a maximum of 60 seconds was reached (90s for 2 month olds). In the test phase, infants saw six further trials, three with the familiar structured grammar and three with with random sequence. The trials alternated with the order counterbalanced. How the sequences were generated in Kirkham et al. (2002) is schematically illustrated in Figure 2.1. In the grammatical or structured case (habituation and test) the stimuli formed a sequence of randomly occurring pairs. In this condition, stimuli were randomly assigned to three sets of pairs for the duration of the experiment with that infant and a sequence was formed from a random succession of these pairs. So that, for example for one infant the turquoise square was always followed by yellow circle, the blue cross by a red octagon and the pink diamond by a green triangle. In the random condition, the shapes were presented in a different random sequence each time, constrained so that each shape appeared with equal probability but no shape appeared twice in a row. Thus across the two conditions the frequencies of the individual shapes were balanced.
2.1. INTRODUCTION

Figure 2.1: The relation between successive shapes in Kirkham et al. (2002).

(Each shape occurring about $1/6$ th of the time). A further illustration is given in Figure 2.15.

Kirkham et al.’s results are summarised in Figure 2.2. All age groups were able to discriminate between the familiar statistical sequences and novel random ones, looking longer at the random sequences. Kirkham et al. take this as evidence for a domain general statistical learning mechanism that is in operation from a very young age. Although looking times are shorter for older infants, Kirkham et al. do not find any developmental differences in discrimination. Although there was a main effect of age, there was no interaction between age and discrimination. It is a well established finding that younger infants look longer overall in looking time paradigms (Fantz, 1964; Johnson, 1996 although note that infants ability to sustain attention increases over time, see Richards and Casey, 1992). Kirkham et al. do not report any data from the habituation phase so it is unknown what if any differences there were in initial exposure to the structured sequence. Whilst this
study did demonstrate that 2, 5, & 8 month olds prefer a more randomly ordered sequence after habituation to a statistically constrained pattern, Kirkham et al. did not run a control condition to see if infants have a prior preference for a less predictable sequence. There are several reasons to suspect that infants may look longer at a random sequence, the total number of possible item transitions is much lower and the structured sequence will be far more likely to have highly regular runs of alternating items. Thus infants could be responding to these low level statistical properties. One aim of the current study is to test for this possibility. We hypothesise that infants will have a spontaneous prior preference for the more ‘random’ sequence. Therefore, Experiment 1 follows as closely as possible the methodology of Kirkham et al. but omits the habituation phase. Additionally, we shall conduct analyses to try to determine whether measures of frequency, complexity or redundancy can provide predictors of infant preference (see next section). Two further experiments will adjust the relative complexity of the two sequences. Finally, we seek to develop an suitable control task to assess adults response to these low level statistical properties in visual sequences. However we begin by surveying various measures of sequence complexity.
2.1.4. Measuring Complexity. We hypothesise that infants will be sensitive to the complexity of the stimuli they are shown. Intuitively this idea has appeal. We can expect an infant to get bored more quickly when viewing a static stimulus or a highly repetitive sequence as compared to a stimulus that is highly dynamic or a constantly changing unpredictable sequence. Although this hypothesis has never been directly tested for sequential stimuli there is evidence that infants respond to stimulus complexity in static displays. In Les Cohen’s seminal study on attention getting and attention holding (Cohen, 1972), 4 month old infants’ attention was held for longer by more complex stimuli which in that case were static checkerboards with greater numbers of squares. This, in turn, was a replication of an earlier finding by Brennan et al. (1966). There was some debate at the time as to whether sustained attention would increase linearly with complexity or would follow an inverted-U shaped profile (Cohen, 1972; Horowitz, 1972). But the issue was not resolved and later research ignored complexity as a predictive measure of infant preference. One reason for this might be due to the lack of a suitable measure of complexity. For such an idea to have any predictive power we need some means of quantifying the complexity of a given set of stimuli. In this section we review several possible measures that might be used to quantify the relative of complexity of sequences of items. Throughout this section we shall consider the complexity of the original paired and random sequences in Kirkham et al. (2002) together with a sequence made up from triplets of items and a completely deterministic sequence of 6 repeating items because these will form the sequence patterns for later experiments in the series. Examples of these are given in Figure 2.15.

2.1.4.1. Item and transition counts. One simple measure of complexity is simply to count the total number of distinct items or of the possible pairwise transitions in a given sequence (square followed by circle, etc.) Kirkham et al. (2002) used the same six items in both the structured and the random sequence but the numbers of possible transitions varied dramatically. For the random sequence each of the six shapes can be followed by any of the five others, giving a total of 6x5 = 30 different transitions (Note that these are all possible non-identical pairings.) For sequences composed of 3 different pairs of items, there are the 3 distinct transitions within the pairs plus 3x3 transitions between pairs giving a total of 12 different transitions. Already this measure gives a large difference between the conditions. Similarly we
may consider the other sequence types show in Figure 2.15. For a sequence built from 2 triplets of items there are 4 within block transitions (2 per triplet) plus 2x2 between block pairs, 8 in total. For a totally deterministic sequence of 6 items repeating in order, there are just 6 transitions.

2.1.4.2. Entropy. However, simply counting transitions does not take account of the relative frequencies or transition probabilities for each pair. Perhaps, a better measure would use the Shannon entropy formula (2.7) first published in Shannon and Weaver (1949).

\[
H(A) = - \sum_{a \in A} p(a) \log_2 p(a)
\]

This is a formula for computing the total 'information' in a given message or string. It is a mathematically rigorous way of quantifying the intuitive idea that the less predictable a given sequence is, the more bits of information it takes to describe it. It sums the probabilities over all possible items in a sequence. For the sequences in these experiments, we know the transition probabilities for all items so we can calculate the entropy measures directly. In fact, we can derive a general measure for each sequence of length \(K+1\). Note that all sequences start with one of the six shapes with equal probability and then follow the transition rules for that sequence type. Note also that \(\log(1) = 0\) so that the predictable/deterministic steps in each case do not add to the entropy.

\[
H(First\ Item) = -6 \left( \frac{1}{6} \right) \log_2 \left( \frac{1}{6} \right) = 2.59..
\]

\[
H_{\text{Random}}(K + 1) = 2.59 - 5.K \left( \frac{1}{5} \right) \log_2 \left( \frac{1}{5} \right) \approx 2.59 + 2.32K
\]

\[
H_{\text{Pairs}}(K + 1) = 2.59 - \frac{K}{2} \cdot 3 \left( 1 \log \left( 1 \right) + \frac{1}{3} \log \left( \frac{1}{3} \right) \right) \approx 2.59 + 0.79K
\]
Figure 2.3: The increasing entropy of four sequence types with sequence length.

\begin{equation}
H_{\text{Triplets}}(K+1) = 2.59 - \frac{K}{3} \cdot 2 \left( 1 \cdot \log(1) + 1 \cdot \log(1) + \left( \frac{1}{2} \right) \log \left( \frac{1}{2} \right) \right) \approx 2.59 + 0.33K
\end{equation}

\begin{equation}
H_{\text{Deterministic}}(K+1) = 2.59 - (K - 1) \cdot 5 \cdot (1 \cdot \log(1)) \approx 2.59
\end{equation}

Plotting these expressions in Figure 2.3 shows how the total entropy increases with sequence length but does so at different rates for the various sequences. It also shows how the difference between the sequences increase with longer sequences. This fits with the intuitive idea that longer exposure improves discrimination and can make some predictions about how easy or difficult it is to discriminate certain sequences. However, as a model of sequence learning, entropy is not psychologically convincing. It takes no account of primacy or recency effects which are known to affect sequence learning (Endress et al., 2005). It does not factor in cognitive load.
or memory effects. For example, all deterministic sequences have constant entropy but clearly longer chains are harder to learn than shorter ones.

2.1.4.3. Item and pairwise frequency. Another online measure that we can compute for a growing sequence is the relative frequency of occurrence of individual items. Although the overall frequency of items in all grammars considered is the same with each item being equally likely, the individual sequences shown to each infant will vary because of the random/stochastic nature of the grammars. Thus at any point in the sequence we can calculate the relative frequency of the last occurring item by dividing its observed frequency by its expected frequency. If infants are primarily sensitive to frequency we may find that they are more likely to disengage from the sequence when they encounter a relatively frequent item that they have seen many times before. A similar measure can be calculated for the frequencies of given pairwise transitions.

2.1.4.4. Local Redundancy. While infants may be sensitive to the global complexity of a sequence, an alternative is that they look away when there is local redundancy (e.g., infants may be more likely to look away when they see a run of items that went \ldots-blue-red-blue-red-blue-red). Repetitive intervals are more likely in constrained grammars and therefore on average infants may look less at these due to more local redundancy in these sequences. Therefore, it is worth exploring measures that could be applied to a moving window within a given sequence to see if the infants look away when these measures are at high values.

As part of an analysis of an implicit learning task in which adults had to recall sequences of coloured discs, Jamieson and Mewhort (2005) introduce a set of formulae for calculating the local redundancy in a given string of N items. Equation 2.13 shows how to calculate ‘zero order’ or single item redundancy score within a given set of N items. Here \(k_i\) is a count of how many times each item \(i\) from the set of all items \(A\) appears in the window.

\[
L_0 = 1 - \frac{1}{\prod_{i \in A} k_i!}
\]
The score of zero indicates that there is no redundancy, namely that all items are
different from each other, whilst a score approaches one as more and more similar
items are encountered, reaching a maximum when all items are the same as each
other. The intuition of the right-most term in Equation 2.13 is that a set of k
completely distinct items can rearranged into k! (k-factorial) different sequences,
whereas however one rearranges a set of k completely identical items there is only
one possible sequence of these. Thus each repeated item reduces the number of
possible ways one can rearrange that set and so increases its redundancy. Taking
this value away from one is just a means of giving this measure in terms of increasing
redundancy.

This formula can be adapted for looking at redundancy at higher orders, in particu-
lar for looking at the 'first order' or pairwise redundancy, that is how repetitive are
the pairs of items in the sequence. This formula is given in Equation 2.14. Here \( k_{ij} \)
is a count of how many items \( i \) & \( j \) appear one after the other in the given window.

\[
L_1 = 1 - \frac{1}{\prod k_{ij}!}
\]

However, there is a short-coming with this measure that leads us to introduce two
new alternative but closely related measures \( R_0 \) and \( R_1 \) given in Equations 2.15
and 2.16 respectively.

\[
R_0 = \sum_{i \in A} \log (k_i!)
\]

\[
R_1 = \sum \log (k_{ij}!)
\]

Figure 2.4 shows how the two redundancy measures compare for all possible combi-
nations a six distinct items with window of size six and indicates why we prefer this
new measure. As can be seen from this graph, the Jamieson and Mewhort (2005)
measure rapidly approaches a limit of one as more items are identical to each other.
Thus as a measure of psychological discrimination it would fail to capture what
would very likely be some highly perceptually salient differences. For example, it
Figure 2.4: Comparing the two redundancy measures for a window of six items. The labels along the horizontal axis show all the possible types of combinations of six items arranged by relative redundancy whilst the vertical axis shows the score given by equations 2.13 and 2.15 respectively. As can be seen from the graph, our newly introduced measure gives more equal steps between the different types of possible sequence, particularly when the sequences have a high degree of redundancy. These all get scores close to 1 with the Jamieson and Mewhort (2005) measure, thus this measure fails to capture the psychological salience in differences between these types of sequence. See text for more detailed explanation.

is highly probable that a participant would notice the difference between various possible runs of six items shown on the right of this graph. There would be a perceptual awareness of the difference between a run of six identical items (e.g. AAAAAA or BBBBBB) and a run containing a second single alternative item (e.g. AABAAA or ABAAAA.) Our new measure R\text{0} captures this fact by spacing each possibility more equidistantly. A further advantage of our measure is that it is not scaled to remain between 0 and 1 irrespective of the size of the window but grows with the sequence length. (This was not a problem in Jamieson and Mewhort’s study where they only considered sequences of a fixed length of 8 items.) Because of the factorial term both measures become less useful when considering much larger window sizes but since we are using the window as a proxy for some kind of short-term memory process we would be unlikely to consider such cases and may as an alternative utilize a more explicit model of memory storage and decay.

2.1.4.5. Explicit models. An alternative quantitative approach to sequential data would be to implement some kind associative model that can learn to find
patterns and structure in sequential data. Neural networks with recurrent connections are an obvious class of such models that can and have been used in this context. Elman (1990) provides a good example of using a simple recurrent network to learn non-adjacent dependencies in a simplified natural language corpus. Whilst Mozer (1993) provides a useful taxonomy of different types of neural network architecture suitable for processing sequential data, in particular emphasising how distinct classes of different memory decay functions can interact at different time scales to capture the contrasts between long and short-term memory and different types of learning. However, for reasons of space and because the internal workings of such networks can often be quite opaque to psychological interpretation, we will not present any neural network models of these results. But we do observe that the data collected is amenable to direct comparison with a model.

2.2. Experiment 1 - Random vs. Pair-based Sequences

Our first experiment aims to investigate infants’ spontaneous online preferences when faced with one random and one statistically constrained sequence without prior exposure. Therefore, we presented infants with the test phase of Kirkham et al. (2002) without any prior exposure. We choose five month old infants because these were the middle group in Kirkham et al. (2002). However, unlike Kirkham et al., we allowed each trial to last up to 90 seconds to allow a great opportunity for discrimination. In all other respects we matched the procedure of the original study.

2.2.1. Method.

2.2.1.1. Participants. The participants were full-term infants who were 5 months (+/- 1 week) at the time of testing and who had experienced no birth complications. Additionally, the infants were screened for familial colour-blindness. Twenty infants (10 female) with a mean age of 154 days (range 144 - 162) were tested. A further 11 infants were tested but are excluded from the analysis because they did not provide full data due to fussiness (5), equipment failure (3), experimenter error (2) or parental interference (1). Infants were recruited via the CBCD’s participant database. Parents were not paid for their infants participation but infants were given a certificate, a small gift and travel expenses were reimbursed.
2.2. EXPERIMENT 1 - RANDOM VS. PAIR-BASED SEQUENCES

2.2.1.2. **Apparatus.** A Macintosh G4 computer running MATLAB (version 5) was used to control the experiment. It was connected to 49cm colour monitor to display the stimuli and a Horita video-titler to superimpose time-synced information on a video of the experiment. Video of the infant was recorded using a low-light DV camera positioned below the screen, facing the infant. These were situated in separate sound-proofed testing booth.

2.2.1.3. **Stimuli.** The stimuli were six coloured shapes (turquoise square, blue cross, yellow circle, pink diamond, green triangle, red octagon) presented in sequence on a black background. To match the method of Kirkham et al. (2002) and to keep the infants attention on the stimuli, the shapes loomed on the screen, increasing from 4cm to 24 cm in height (2.4 – 14.6°) over the course of a second. Shapes were presented one after the other with no interval between them. Each sequence was presented for as long as the infant attended to the monitor. An observer viewed the infant on a separate monitor in the main lab and started and finished each trial on the basis the infants’ attention. The observer held a key down on the computer whenever the infant was attending, releasing it if he or she looked away. The trial was finished if the computer determined that the infant had looked away continuously for 2s. or if the trial had lasted for a total of 90s. Trials always finished when the current shape had finished looming. In the majority of cases the observer was unaware of which type of sequence the infant was observing. However, this blind scoring variable was included the subsequent analysis to see if it had any effect. Between trials the screen showed a small blinking fixation square and infants attention was attracted back to the monitor by the experimenter speaking through the microphone.

2.2.1.4. **Procedure.** Infants were seated in a car-seat, in a darkened room, at a distance of 95cm from the screen. The care-giver remained in the room but was seated behind and out of view of the infant and was instructed not to interact with the infant.

In the random condition, the shapes were presented in a different random sequence each time, constrained so that each shape appeared with equal probability but no shape appeared twice in a row. In the structured condition, stimuli were randomly assigned to three sets of pairs for the duration of the experiment with that infant
and these pairs formed a sequence by a random succession of these pairs. Thus for one infant, turquoise square was always followed by blue cross, yellow circle by a pink diamond and green triangle by a red octagon, whereas for another infant the fixed pairs would be different. This is schematically illustrated in Figure 2.1 and in Sub-figures 2.15(a) and 2.15(b).

Each infant saw a total of six trials, three structured and three random. Presentation alternated between structured and random trails. Ordering of trials was counterbalanced across infants so that half the infants saw the structured sequence first, the other half saw the random first.

2.2.2. Results. Twenty infants were tested. From the video recordings of the experimental trials an experimenter generated two sets of data for the analyses. Firstly was a measure of total looking time per trial for each infant, measured by adding up all the time that the infants were looking at the screen in each trial. Secondly, the exact sequence of shapes seen on a given trial allowing for glances away was produced. Infants were deemed not to have seen a particular stimulus if they saw less than either the first 300ms or last 100ms of that stimulus. A second experimenter, blind to the hypothesis, double coded a randomly selected 25% of the data and results gave a Pearson correlation of \( r = 0.981, p < 0.001 \) with the original coder.

2.2.2.1. Looking time analyses. For the purposes of analysis, the 6 trials were divided into three blocks (block 1 = trial 1 & 2, block 2 = trials 3 & 4, block 3 = trials 5 & 6) each consisting of a pair of sequences; one structured, one random. Thus there are the two within subjects measures (Block and Pattern) in all subsequent analysis. The order in which the infants saw the sequences (structured first or random first) was a between subject variable. Thirteen out of 20 infants were tested by an experimenter blind to which type of sequence the infants were viewing on a given trial. A two-within, two-between mixed ANOVA (\( 3 \times 2 \times 2 \times 2 \), Block x Pattern x Order x Blind) showed no significant main effect of blind-scoring \( F(1,16) = 0.65 \) nor any significant interactions with other factors (all \( F \)'s < 1). Therefore the scoring was considered equivalent and all data were pooled across this factor for subsequent analysis.
2.2. EXPERIMENT 1 - RANDOM VS. PAIR-BASED SEQUENCES

Figure 2.5: Graph of overall results for Experiment 1. Mean looking times (in seconds) for random and structured colour and shape sequences for 5 month old infants. There was a decrease in looking over the three pairs of trials and infants consistently looked longer at the random sequence. Error bars are 95% confidence intervals.

To investigate whether the type of sequence had an influence on the length of looking, a two-within, two-between mixed ANOVA (3 x 2 x 2 x 2, Block x Pattern x Order x Gender) was performed. Figure 2.5 summarises the results. As expected there was a highly significant main effect of block, F(2,32) = 13.2, p<0.001. Infants looking times decreased across the three pairs of sequence presentations indicating that they became familiar with the overall experimental set-up. There was no significant main effect nor any interactions between order and the other factors, indicating that the order of presentation of random and statistical sequences did not have any effect on infants’ discrimination of these cases. In line with our hypothesis, there was a significant main effect of pattern, F(1,16) = 5.49, p<0.032. Infants looked longer on average at the random pattern in each presentation trial. However, the effect of pattern is qualified by an unexpected effect of gender in this sample. There was a significant interaction between gender and pattern F(1,16) = 5.17, p<0.037, the male infants showed a strong and consistent preference for the random sequence, looking longer on average at each of the three pairs of presentation. In the absence of any strong preference in the female infants, this male effect...
largely accounts for the significant main effect of Pattern. There was a highly significant main effect of Gender $F(1, 16) = 10.2$, $p < 0.006$, male infants looked longer overall than females. These results are summarised in figure 2.6. One possible reason that males showed the hypothesised preference for random over statistical sequences could be related to their longer overall looking times. Longer exposure to the sequences provides more exemplars from which the difference in complexity of the two sequences could be detected.

To investigate if longer looking alone is driving discrimination, a further analysis of the data in Experiment 1 was carried out, dividing the infants by a median split on total looking time into a long-looking and short-looking group. Seven of the ten longer looking infants were male, so there is a potential confound between these two explanations. However, if as we speculate above, it appears that longer looking leads to great discrimination then this could provide a more parsimonious explanation of the gender interaction, although an explanation would still need to be found for why male infants are looking longer. Figure 2.7 shows the data grouped in this fashion and we also performed a two-within, two-between ANOVA (3x2x2x2,
Block x Pattern x Order x Long/Short Looking). As expected this manipulation gave a highly significant main effect of long/short total looking $F(1,16)=27.24$, $p<0.001$ and a large main effect of Block $F(2,32)=20.8$, $p<0.001$. There was a highly significant interaction between Block and long/short looking time $F(2,32)=10.5$, $p<0.001$. The long looking group showing steep decline in looking over the three blocks, whilst the short looking did not. The main effect of pattern remained, $F(1,16)=6.38$, $p<0.022$, namely infants looked longer at the random sequences than the structured sequences. However, there were no significant interactions between pattern and the long/short looking variable. The expectation being that if detecting the pattern was mediated purely by looking time then there would an interaction between these two factors. No such interaction was found suggesting that some kind of gender difference is a better explanatory factor for discrimination than just the difference between long lookers and short lookers, although the fact that male infants do look longer is likely to partly account for their greater discrimination.

2.2.2. Complexity analyses. As outlined in section 2.1.4 another way to look at the data is from the perspective of the relative complexity of the particular sequences. For any given measure we might postulate that if infants were sensitive to it then they be more likely to disengage from the sequence when that measure
takes an extreme value. For example, if infants are sensitive to item frequency, they may look away when they see a shape they have seen a large number of times before and although these frequencies are balanced overall, the random nature of the particular sequences used will mean there is always local variation. Likewise, local measures of redundancy will vary throughout the trials and on a given trial infants may look away as a result of particularly repetitive sub-sequence. Notice also that repetitive sub-sequences will be more likely in a structured grammar leading to a potential confound between infants learning global aspects of the grammar and infants looking away sooner because of local redundancy.

Therefore, to see if local redundancy or item frequency can account for infants looking away we examined how these measures change up to the end of a given trial. If an infants sensitivity to a certain type of redundancy drives their looking away then we would expect that measure to be higher than average at the end of a given trial. Therefore we look at the redundancy scores at each point leading up to the end of a given trial and compare this to the overall average. This is illustrated in Figure 2.8 for $R_0$ measure with a window size of 8 items. (See Section 2.1.4.4 for explanation of how $R_0$ and $R_1$ are calculated.) We calculate these scores at each point in the sequence (looking back on the previous 8 items) to get an overall average per trial and we then compare this to the same score at the point of presentation of each of the last 6 items. If this measure is related to the infant looking away we would expect this curve to have an upwards slope and at least the final value to be significantly higher (i.e. greater redundancy) than the average. A one-tailed, paired-sample t-test comparing the final $R_0$ score with the average $R_0$ across that trials showed that this value was significant, $t=2.51$, d.f. = 105, $p<0.007$. The same analyses were conducted for $R_0$ with a six item window and $R_1$ with 6 and 8 item windows. All were significant at the 0.05 level. These results are summarized in Table 2.1. Note that all of these redundancy measures will be somewhat correlated (especially with large windows) so we did not look at window sizes of greater than 8 items. The same analyses with a window size of 4 items were not significant. This suggests that infants are sensitive to both item and pairwise redundancy, and look away when the sequences are become more locally redundant.

We also conducted two analyses to see if the global relative frequency of individual items and pairs of items was related to infants looking away. This is a cumulative
2.2. EXPERIMENT 1 - RANDOM VS. PAIR-BASED SEQUENCES

Figure 2.8: Graph of the zero order redundancy scores calculated using Equation 2.15 with a window size 8. See text for explanation.

Table 2.1: Redundancy and relative frequency measures for Experiment 1. See text for explanation.

<table>
<thead>
<tr>
<th>Measure (window size)</th>
<th>Average per trial</th>
<th>redundancy and frequency scores</th>
<th>T-tests*</th>
<th>p(last &gt; average)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0(8)</td>
<td>2.44</td>
<td>2.47 2.48 2.42 2.52 2.57 2.68</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>R0(5)</td>
<td>1.24</td>
<td>1.25 1.28 1.25 1.29 1.32</td>
<td>0.029</td>
<td></td>
</tr>
<tr>
<td>R0(4)</td>
<td>0.40</td>
<td>0.46 0.50 0.50 0.53 0.59</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>R1(8)</td>
<td>0.94</td>
<td>1.04 1.04 0.95 0.95 0.95</td>
<td>0.039</td>
<td></td>
</tr>
<tr>
<td>R1(6)</td>
<td>0.45</td>
<td>0.47 0.47 0.46 0.50 0.53</td>
<td>0.035</td>
<td></td>
</tr>
<tr>
<td>R1(4)</td>
<td>0.15</td>
<td>0.18 0.18 0.19 0.19 0.18</td>
<td>0.159</td>
<td></td>
</tr>
<tr>
<td>Prec(items)</td>
<td>0.03</td>
<td>0.05 0.05 0.05 0.04 0.04</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Prec(items)*</td>
<td>0.26</td>
<td>0.27 0.27 0.25 0.22 0.18</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

* - zero order alternative redundancy measure
R1 - first order alternative redundancy measure
* - 1-tailed repeated measures t-tests
† - arithmetic mean of log frequency scores

Table 2.1: Redundancy and relative frequency measures for Experiment 1. See text for explanation.

function and the calculation is somewhat involved so we explain it here. To find the relative frequency of a given item (or pair) in the sequence one needs to look back over everything that has been seen so far and count the total number of occurrences of that item, this gives an actual frequency of that item (pair). To get a relative frequency we divide this number by the unbiased expectation of how frequent that item (pair) has been. For example, assume we are at the end of trial 1 and the infant has seen 60 shapes in total and the last shape is a square. Given the six possible shapes we would expect the square to have appeared \( \frac{1}{6} \times 60 = 10 \) times on average (this is the unbiased estimate). If this is in fact the 15th square then it will have a relative frequency of 1.5. Now, later in the experiment at the end of trial 3 where the infant has seen a total of 120 items and the last item, a circle has appeared 17 times previously then it has a relative frequency of \( \frac{17}{20} = 0.85 \). For each infant we can calculate these values at each point in the experiment and
2.2. EXPERIMENT 1 - RANDOM VS. PAIR-BASED SEQUENCES

Look at the values in the last few steps of each trial. As with the redundancy scores, we would expect the relative frequency of the last items to be greater if the infant looks away when seeing a globally frequent item. However, in this case we must be careful what we compare this value with. Firstly, it is not appropriate to take an arithmetic mean of relative frequencies so we log transform them first. Secondly, because the global relative frequency is a continually changing measure throughout the experiment, it is not appropriate to take a global average of this value. The best we can do is to compare the relative frequency of the last item or pair seen with its immediate predecessors. Graphs of these two measures (for single items and pairs) are shown in Figure 2.9 and the actual values are given in Table 2.1. As can be seen, neither measure appeared to be predictive of the end of the trials, since the relative frequency of the last item does not differ from that of its predecessor. This suggests that infants do not look away on the basis of globally frequent individual items or pairs.

2.2.3. Discussion. In Experiment 1 we found that 5 month old infants spontaneously preferred (looked longer at) a random sequence compared to a more predictable sequence (paired items) of the same shapes. Frequencies of individual
items were equal, so infants can only discriminate on a basis of pairwise transitions or some measure of sequence complexity. They discriminate spontaneously from the first block without any prior exposure to either sequence type. This suggests that a process simpler than artificial grammar learning is in operation. Potentially, they are sensitive to the relative complexity of the sequences. Results from complexity analyses seem to support this conclusion. We found infants looked away from the sequences when they exhibited greater zero and first order local redundancy. But that the relative frequency of items or pairs was not greater at the point of look away, suggesting that infants do not track overall item or pairwise frequencies.

There was an unexpected main effect of gender with male infants looking longer overall. Additionally there was an interaction of pattern-discrimination with gender with male infants discriminating better than female infants. One possible explanation for this better male discrimination could paradoxically be because the female are habituating faster. Using the same task as Cohen (1972), both Caron and Caron (1969) and Creighton (1984) found that at 4 months old female infants habituated faster than male infants. This could be the explanation for the male infants overall longer looking times which in turn this longer looking could explain their greater discrimination. Infants who look longer in total will accumulate more examples and thus have more information about the differences between the two conditions. However, this faster female habituation is not commonly reported in other tasks and indeed Cohen (1972) found mixed pattern of responding which if anything showed make infants habituating faster. Alternatively, it could be that males are relatively more interested in abstract geometric shapes and look longer to stimuli of this kind (Kimura, 1999). However, it is not known for certain why male infants would exhibit longer looking times. Indeed, very few studies have reported any sex differences in infants cognitive abilities (Spelke, 2005).

These results suggest a number of directions to follow. Firstly, finding the effect of pattern indicates that the infants may possess some prior preference for more complex stimuli and running a task with a different grammar will allow an investigation of this effect. Additionally, we wish to see if the unexpected sex differences persist in a second task. Experiment 2 below investigates these possibilities. Using sequences built from three pairs or two triplets of items, it would be expected on the basis of information theory that the former being more complex would hold
infants' attention for longer. Thus, we would expect a similar pattern of results to Experiment 1.

2.3. Experiment 2 - Pairs vs. Triplets, High Overlap

It seems from Experiment 1 that even in the first pair of trials, the infants are already discriminating between the random and structured sequences. This could be interpreted as a sensitivity to the differing complexity of the two sequences. Alternatively, they could have very rapidly learned some aspects of the grammar, perhaps a single pairing of shapes, the repetition of which leads them to disengage from that particular sequence. These two hypotheses are not mutually exclusive, since grammar learning involves learning of the statistical properties. If the infants are responding directly to the overall complexity of the sequence rather than the recognition of aspects of the grammar then we would predict that in a new task where both sequences have a grammatical structure but the difference in complexity between two sequence was comparable to that found in experiment 1, the infants would again look longer at the more complex sequence.

This is the aim of Experiment 2. In this case we compare the structured sequence from Experiment 1 (made up of the three pairs of shapes) to a more predictable sequences consisting of the same shapes constrained to always be in one of two triplets. If infants look longer at more complex sequences then in this case they should look longer at the pair based sequence compared to that based on triplets.

2.3.1. Method.

2.3.1.1. Participants. Twenty full-term infants (10 female) were tested with a mean age of 153 days (range 146 - 164). A further 5 infants were excluded due to fussiness (4) or parental interference (1). Infants were screened and recruited as in Experiment 1.

2.3.1.2. Procedure & Stimuli. The procedure was identical to that in Experiment 1. Each infant saw a total of six trials, the length of each one controlled by the infant's looking at the screen. Each trial ended when the infant looked away for 2 seconds or reached a maximum of 90 seconds. In all cases, this was determined by the computer controlled by an experimenter blind to the trial type. The
trials alternated between 3 sequences consisting of (randomly selected) pairs and 3 sequences consisting of triplets. The same six shapes as in the previous experiment were used in both cases. Figure 2.15 helps illustrate the types of sequence used in this experiment.

2.3.2. Results. The videos were coded as before and second experimenter blind to the hypothesis double coded a randomly selected 25% of the data. The second coders results gave a Pearson correlation of $r=0.996$, $p<0.001$ with the original coder.

2.3.2.1. Looking time analyses. As before, for the purposes of analysis, the 6 trials were divided into three blocks each consisting of a pair of sequences; one structured, one random. Thus, there were two within subjects measures (Block and Pattern) in all subsequent analysis. A two-within, two-between mixed ANOVA (3 x 2 x 2 x 2, Block x Pattern x Order x Gender) was performed. Figure 2.10 summarizes the results. As before, there a highly significant main effect of Block, $F(2,32)=8.28$, $p<0.001$. The infants looked less at later trials than earlier ones. There was only a trend towards boys looking longer than girls and no interaction between gender and other factors. Contrary to the original hypothesis there was no main effect of pattern, infants did not consistently look longer at the more complex pair-based sequence. Although a post-hoc test confirmed that they did do so in the first block, $t=1.89$, d.f.=20, $p<0.037$.

There were unexpected interaction of pattern with order, $F(1,16)=9.64$, $p<0.007$ and a marginal interaction between pattern and block, $F(2,32)=3.07$, $p<0.06$. Order refers to which pattern infants saw first; 9 infants saw the pairs sequence on 1st, 3rd & 5th trials and the triplets sequence on 2nd, 4th & 6th trials, whilst 11 infants saw triplets then pairs. Figure 2.11 shows the infants grouped by order of sequences with the trials in the exact order they were seen. From this, it can be seen that overall there is no interpretation consistent with infants looking longer a just one type of sequence.

2.3.2.2. Methodological Problem with Experiment 2. The analysis suggests that infants are not discriminating sequences of pairs from sequences of triplets in this experiment. At least, not as clearly as in Experiment 1. Looking again at the data, a mistake in the presentation code was discovered. For each infant, the groups of
2.3. EXPERIMENT 2- PAIRS VS. TRIPLETS, HIGH OVERLAP

Figure 2.10: Graph of mean looking times in Experiment 2 to pair and triplet based sequences grouped by gender.

Figure 2.11: Graph of mean looking times in Experiment 2 to pair and triplet based sequences grouped by sequence order with trials in exact order they were shown.
2.3. EXPERIMENT 2- PAIRS VS. TRIPLETS, HIGH OVERLAP

Figure 2.12: Example of the type of sequences used in Experiment 2. Note that a procedural error lead to the two types of sequence having maximal grammatical overlap.

shapes which formed the pairs and triplets were supposed to be randomised. Although this was done once per infant to get the initial mappings circle = A, square = B, etc to form the triplets \{ABC\}, \{DEF\}, the same mappings were used to produce the pairs \{A'B'\}, \{C'D'\}, \{E'F'\}. These two sets are maximally overlapping in terms of the transitions they contain. There are 6 pairs which occur in both cases. And in particular, the four transitions having probability 1 in the triplets case will all be present in the pairs sequences. This can be seen in the example given in Figure 2.12.

With these highly overlapping stimuli sets, we should not be too surprised if there is very little discrimination, especially in later trials. Once the infant has learnt something about either sequence then some of this knowledge will transfer to the pairings seen in the other case. Thus overall we might expect a quick habituation and less attention in later trials. This appears to be the case. Furthermore, while the similarities between the two sequences could explain the lack of discrimination in later trials, the asymmetries between them could explain the interactions seen in trials 1 & 2 (the left-most data in Figure 2.11) because the sequences share 6 common transition pairs, there is an asymmetry moving from one to the other.

The pairs sequence has more variety overall so we may expect longer spontaneous looking in the very first trial. Seeing the triplets sequence following this only introduces 2 novel pairings so we would expect much reduced looking. By contrast, seeing triplets first followed by pairs, starts with a relatively less complex set and introduces 6 novel pairs on trial 2. Performing a repeated measures analysis of variance (Trial x Condition) on just the first block we see that the interaction is significant \(F(1,18)=6.24, p<0.022\).

Because of the methodological error in this experiment, we do not perform any complexity analyses.
2.3.2.3. Comparing the results of Experiments 1 and 2. Further analysis of the data would be possible with experiment as a between subjects condition. Given that we did not find any strong overall discrimination between sequence type in Experiment 2, it becomes difficult to directly compare Experiments 1 & 2, trial by trial. Instead we can look at the overall looking time to the more and less structured patterns between experiments and between genders with a 1 within-subject, 2 between ANOVA (2x2x2, Pattern x Experiment x Gender). Here we find a highly significant main effect of gender, $F(1,36) = 9.90, p<0.003$. Male infants are consistently looking longer overall. There is a marginal main effect of pattern $F(1,36) = 3.78, p<0.06$. But there are no other significant effects or interactions. This suggests that the effect of complexity on looking time persists across the two experiments even given the weak findings of Experiment 2.

2.3.3. Discussion. Due to a methodological error, Experiment 2 presented a much more difficult discrimination than intended, presenting two grammars with a high degree of overlap. As well as having the same individual shapes at same frequencies, the sequences had 6 transitions in common. Now the infants showed a significant preference for the more complex sequence (pairs) only in the first block. There was also an interaction between order of presentation and the discrimination. These results were nevertheless consistent with the idea that infants have spontaneous sensitivity to complexity. Moreover, these results were consistent with the idea that infants were learning features of grammars presented. If we consider the first block of trials shown in Figure 2.11, the asymmetry between the item-pairings in the triplet and pairs sequences could explain the interaction seen in this block of trials. Seeing first pairs then triplets introduces only 2 novel pairings and so leads to a large drop in interest, conversely seeing triplets then pairs starts with a less complex set and introduces greater novelty on the second trial holding infants’ interest. Finally, there was a trend towards male infants looking longer. To clarify exactly what might be happening we re-ran Experiment 2 using grammar with much lower degree of overlap.
2.4. Experiment 3 - Pairs vs. Triplets, Low Overlap

Experiment 3 repeats the approach of Experiment 2 but using pairs and triplets specifically chosen to have minimal overlap in the pairwise transitions. We choose the pairs and triplets such that there are only two pairwise transitions that are common to both sequences. An example is shown in Figure 2.15 (b) and (c) where only the transitions {yellow circle -> blue cross} & {red octagon -> cyan square} are common to both grammars.

2.4.1. Method.

2.4.1.1. Participants. A total of 17 infants (8 female) completed the experiment, their mean age was 154 days (range 137 -163). A further two were excluded due to experimental error and equipment failure, whilst 8 were excluded due to fussiness, 6 of these 8 had already taken part in a previous unrelated experiment which may have accounted for their fatigue. Three of the included infants also took part in a previous experiment but there was no indication that they were fatigued and their data were retained. Infants were screened and recruited as in Experiment 1.

2.4.1.2. Procedure & Stimuli. The procedure was identical to Experiment 1. Each infant saw a total of six trials, the length of each one controlled by the infant’s looking at the screen. Each trial ended when the infant looked away for 2 seconds or a maximum of 90 seconds. In all cases, this was determined by an experimenter blind to the trial type. The trials alternated between 3 sequences consisting of (randomly selected) pairs and 3 sequences consisting of triplets. The same six shapes as in experiment were used in both cases. Figure 2.15 (b) and (c) helps illustrate the types of sequence used in this experiment.

2.4.2. Results. The videos were coded as before and a second experimenter blind to the hypothesis double coded a randomly selected 25% of the data and results gave a Pearson correlation of r=0.989, p<0.001 with the original coder.
2.4.2.1. Looking time analyses. As in Experiment 1, we looked for any effects of gender or pattern complexity on the infants looking times. Therefore, we conducted the same repeated measures analysis of variance (Pattern x Block x Order x Gender) with Pattern and Block as in the previous experiments. As before there was a highly significant main effect of Block $F(1,13)=32.8, p<0.001$. There was also a highly significant Pattern by Order interaction $F(1,13)=13.6, p<0.003$. There were no other significant main effects or interactions. The main effect of block was a general decrease in looking throughout the experiment. The interpretation of the Pattern by Order interaction can be seen by examining Figure 2.13 where all the results are plotted in the order in which the trials, grouped by the Pattern condition. Here we can see that the interaction is actually a consequence of there being no main effect of pattern but a general decline in looking over the trials. Thus it appears that infants do not look longer at either the sequences generated by a pair-based or a triplet-based grammar.

2.4.2.2. Complexity Analyses. Again, we are interested in what other factors might influence when the infants look away from the stream of shapes. Therefore an identical set of analyses to Experiment 1 (section 2.2.2.2 ) were carried out.
Table 2.2: Redundancy and relative frequency measures for Experiment 3. See text for explanation.

The results are summarized in Table 2.2. Again the zero-order and first-order redundancy measures, $R_0$ and $R_1$ show the expected pattern, infants appear to look away at points when the sequence is more repetitive and redundant than average. All the measure show at least a trend in this direction. This time the four-item windows also showed the redundancy effects which may reflect the partial learning of some two item pairs. The results of global relative frequency analyses are also presented in Table 2.2 and are shown graphically in Figure 2.14. This time there appears to be evidence for infants looking away when encountering more frequently occurring pairwise transitions.

2.4.3.2. Comparing Experiments 1 and 3. Although Experiment 3 showed no effect of grammatical complexity nor any gender effects, it may still be instructive to compare the results with those of Experiment 1. For example, it may be that global difference in complexity between random and pair-based sequence has some effect in Experiment 1 where the difference is large enough to be noticed, but that the difference between pairs and triplets is too small to be noticed in Experiment 3. In this case, we might possibly still expect to see a between subjects effect with participants in Experiment 1 staying engaged for longer as group than those in Experiment 3. In fact, if we run a comparison on the total looking time per participant we find that the means of these values are almost equal with no significance $t<1$. The mean total looking in Experiment 1 was 199 seconds and in Experiment 3 it was 197 seconds.

2.4.3. Discussion. In Experiment 3 infants were shown sequences generated from two triplets and sequences generated from three pairs of coloured shapes, cho-
2.4. EXPERIMENT 3 - PAIRS VS. TRIPLETS, LOW OVERLAP

Figure 2.14: Graph of the relative frequency scores for single and pairwise items in Experiment 3. See text for explanation.

sen so that there was minimal overlap between the two sequences. It was found that infants looked equally long at both types of grammar. This contradicts the hypothesis that infants would look longer at the globally more complex (pair-based) sequence. This may be because infants are not in fact sensitive to the global complexity of sequences or it may be that the difference between the two sequence types was below the threshold of their awareness. However, a secondary prediction of a global complexity was also not supported; when overall looking times were compared, participants did not attend less overall in Experiment 3 than in Experiment 1.

The results of Experiment 3 were also analysed in terms of whether local redundancy and the relative frequency of shapes and pairs of shapes could predict when infants would look away from the sequences. As in Experiment 1, it was found that single item and pairwise redundancy measures were significantly above average at the point where infants looked away, particularly when considered over a six item window. Infant were more likely to look away when the most recent few items they had seen were more repetitive. Additionally, there was a global effect detected in
the pairwise relative frequency measure, suggesting that infants looked away when they encountered relatively more frequent pairs of items. Taken together these analyses suggest that infants may be learning aspects of the grammar, responding to the more frequently encountered elements. Potentially this supports a chunk-based interpretation.

2.5. Experiment 4 - Adult Control Task

Anecdotally, we found that in all experiments, the parents were generally unable to discriminate between the two types of sequence they were seeing with their infants, even though they had usually been informed of the type of structures that would be presented. This is interesting because it could potentially be an example of rule-use in which infants outperform adult possibly because the infants’ implicit learning abilities are not interfered with by explicit conscious mechanisms. Alternatively, it may be that adults would succeed on the pairs versus random discrimination if tested in the right way and possibly also on the pairs versus triplets task that the infants fail. The current experiment aims to investigate adult abilities.

Clearly a looking time paradigm is not appropriate for use with adult participants. Therefore, we designed a task in which adult participants must determine which of two 30 second presentations of the looming sequences of coloured shapes has more underlying structure (or equivalently is less random). However we do not wish the adults to engage too heavily in an explicit conscious effort to solve the task because then we testing a very different set of abilities to those likely to be at work in the infants. One traditional way to disengage explicit strategies would be to have a dual task but adding a second task would add other extraneous complications of its own. Therefore for this study we settled on using a forced choice guessing procedure where participants are encouraged not to think too hard about the solution of the problem but “to go with what feels right.” To encourage them to take this approach we also had them rate their confidence in their answers, emphasizing that guessing was permitted.

The effect of gender in Experiments 1 and 2 was unexpected. Therefore, we are interested in investigating if any gender effects persist in adults.
2.5.1. Method.

2.5.1.1. Participants. 28 adult volunteers (14 male) were recruited by means of an opportunity sample in and around Birkbeck College. They were informed they would be participating in a study of patterns and randomness and their written consent was obtained before the study took place. They were each paid five pounds for their participation.

2.5.1.2. Apparatus. A Macintosh Powerbook G4 computer running MATLAB R2006a (version. 7.2) was used to display the stimuli and record participants responses. Participants were seated in front of the computer and controlled the display of stimuli and entered their responses by means of the keyboard. All participants were tested in a quiet empty room.

2.5.1.3. Stimuli. The stimuli were the same six coloured shapes used in the previous experiments presented on a black background. The exact order and co-occurrence rules were determined by the computer program according to the test conditions described below. Shapes were presented one after the other with no interval between them at the rate of about 1 per second. They appeared in the centre of the screen looming from about 3cm to 14cm across.

There were four different sequences types, as illustrated in Figure 2.15. Sequence type 1 matched the random condition of Experiment 1, items appeared with equal probability but with no item repeated. Type 2 matched the structured pairs condition of Experiment 1, items were group into three pairs. Items within the pairs always occurred in the same order (p = 1.0 ). Whilst all pairs were equally likely (p=0.333) to follow each other. Sequences of type 3 consisted of two sets of three distinct items. These triplets always appeared together in the same order (p=1.0) but transitions between triplets was random with each one equally likely (p=0.5). As in Experiment 3, the triplets had minimal overlap with the pairs. Finally, sequences of type 4 were completely deterministic. The six items always occurred in the same order and the order looped repeatedly through these six items and again had low overlap with the deterministic transitions of the pair and triplet sequences.

The exact item combinations for types 2,3,4 varied between participants depending upon a random seed entered into the control program. However, the particular
dependencies were constant for all the 12 trials that a given participant saw. Each trial started with a shape selected randomly and then continued following the appropriate transitions for that sequence type.

2.5.1.4. Procedure. Participants were tested individually. At the start of the experiment the participants were invited into the testing room and were seated in front of the computer. The experimenter collected some basic demographic information and confirmed that the participant did not have colour-blindness (as far as they were aware).

The experimenter instructed them that they were going to see pairs of sequences of coloured shapes and their task was for each pair to determine which of the two sequences was more ordered or (equivalently) less random. Their understanding of this concept was checked by means of a simple example involving sequences of heads and tails. Furthermore, the participants were informed that this task was comparing performance to that of infants and therefore they were not required to concentrate too hard on the task rather trusting their first impressions and intuitions. They were instructed that for each pair of items they would have indicate how confident they were about their choice of the more ordered one. They were to do this by giving a rating on scale of 1 to 7, where 1 indicated a complete guess and 7 absolute certainty. When the participants were happy with the instructions, the experimenter left the room and the trails began.

Figure 2.15: Illustration of the transition probabilities in the four types of test sequence. See text for explanation.
Each trial consisted of a two 30 item sequences presented one after the other. Each sequence of one type being paired a sequence of another. With four sequence types, there are therefore 12 possible pairings (1 then 2, 2 then 1, 1 then 3, etc.) Each participant saw a total of 12 trials consisting of all possible pairings. The order of trials was pseudo-randomized across participants. Each trial began when the participant pressed a key on the keyboard to start the first sequence. The 30 shapes in the first sequence appeared and the participant was prompted to press a key to start the second sequence. After the second set of 30 shapes had finished, the participant was prompted to indicate with one they considered to be more ordered or less random by pressing either key '1' or key '2' on the keyboard. Next the computer requested that they give a judgment of how confident they were about their decision on a scale from 1 to 7, also using the appropriate numeric keys.

At the end of the experiment, the experimenter came back into the room and questioned the participants about the task. They were asked if they thought they had spotted any particular patterns and what, if any, strategy they had employed. Including, if they had concentrated more on colour or shape and if they had subvocalized or mentally repeated the names of the objects. The actual structure and pattern of the sequences was then explained to them and any further comments were recorded. Finally, they were paid and thanked for their participation.

2.5.2. Results. A total of 28 participants (14 female) were tested. For each participant we recorded their accuracy on each of the six types of pairing, grouping the answers according to the type of sequences that were being compared (pairs vs random, pairs vs. triplets, etc.) Thus for each participant there were six scores out of 2 (because 1 vs. 3 is the same comparison as 3 vs. 1). The results grouped by sex are summarised in Figure 2.16. For each group and each question type we calculated the probability of obtaining that many correct responses by chance by comparing total score to the binomial distribution with N=28 (14 participants per group each answering two questions.) As can be seen, the pattern of response was largely the same between the sexes with both male and female being at chance on pairs versus triplets (2-3) and triplets versus deterministic (3-4) and on all other comparisons the participants tended to perform statistically better than chance. In
2.5. EXPERIMENT 4 - ADULT CONTROL TASK

Figure 2.16: Bar chart of the accuracy of participants in the adult sequences task. Each column represents the mean correct responses for each of the possible sequence pairs, grouped by gender. For example, the first (pink) column represents the female performance on the task comparing pair sequences (type 2) with triplet sequences (type 3). Each participant has one trial in each order (2 vs 3 and 3 vs 2) so the maximum score is 2, chance performance (1) is indicated by the dotted line. The pairs are arranged in terms of their relative difference in complexity of the sequences. Significance levels based on a binomial distribution, * p < 0.05, ** p < 0.005, *** p < 0.0005.

Table 2.3: The differences in entropy for the 4 sequence types compared in Experiment 4. The first two columns show the two sequence types, the third and forth column show their respective entropy scores whilst the fifth column show the difference. Also shown are the number of distinct transitions in each sequence, which provides a simpler measure of sequence complexity.

<table>
<thead>
<tr>
<th>Seq A</th>
<th>Seq B</th>
<th>H(A)</th>
<th>H(B)</th>
<th>H(A)-H(B)</th>
<th>Trans(A)</th>
<th>Trans(B)</th>
<th>Trans(A) - Trans(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>4.02</td>
<td>1.00</td>
<td>3.02</td>
<td>8</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>10.15</td>
<td>6.82</td>
<td>3.33</td>
<td>12</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>10.15</td>
<td>1.00</td>
<td>9.15</td>
<td>12</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>27.87</td>
<td>10.15</td>
<td>17.72</td>
<td>30</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>27.87</td>
<td>6.82</td>
<td>21.05</td>
<td>30</td>
<td>6</td>
<td>24</td>
</tr>
</tbody>
</table>

terms of overall accuracy women (67% correct) marginally outperformed men (61% correct).

To investigate whether difference in complexity between sequences was a predictor of the participants discrimination ability we used equations (2.9 -2.12) with k = 30 to give an entropy score for each sequence type and took the difference between
Table 2.4: Results of a Logistic Regression Analysis of adults’ ability to discriminate two sequence types. Shown are the parameters for a model based on a measure of difference in entropy between sequences and a parameter for the gender of the participants.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>β</th>
<th>S.E.</th>
<th>Wald's χ²</th>
<th>df</th>
<th>p</th>
<th>Exp(β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H(A) - H(β)</td>
<td>.624</td>
<td>.013</td>
<td>3.347</td>
<td>1</td>
<td>.067</td>
<td>1.824</td>
</tr>
<tr>
<td>Sex (0 = female, 1 = male)</td>
<td>-.235</td>
<td>.229</td>
<td>1.055</td>
<td>1</td>
<td>.304</td>
<td>.799</td>
</tr>
<tr>
<td>Constant</td>
<td>-.357</td>
<td>.244</td>
<td>2.141</td>
<td>1</td>
<td>.149</td>
<td>.288</td>
</tr>
</tbody>
</table>

these for each possible pairing. A summary of these values is shown in Table 2.3 and note that the data in Figure 2.16 are arranged in order of increasing difference in complexity. This score can be used as a predictor variable in a binary logistic analysis for the likelihood of participants correctly determining which sequence in a given pair was more random. Sex was also included as categorical predictor variable since women appeared to outperform men. A summary of the model is found in table 2.4. Overall the model does not account for the data well. Although the model has a marginally significant coefficient for the entropy difference as measured by the Wald’s χ², test, χ²(1) = 5.11, p<0.067, the overall model evaluation was not significant, χ²(2) = 4.43, p<0.11. The Hosmer-Lemeshow test is a goodness of fit measure that compares the proposed model to the actual data, this gave a χ²(8) = 5.11 which was insignificant (p > 0.7), suggesting that the model does at least fit the data. However, it has a very low Nagelkerke R² of 0.018. This a proxy measure for the R² found in regular regression models but which does not have a well-defined meaning in a binary logistic regression, nevertheless a very low value of the Nagelkerke R² indicates that the model is not account well for the data. This suggests that differences in entropy of the sequences was does not account for performance across all sequence comparisons. Additionally, this analysis indicates that there were no significant differences between the sexes.

As highlighted in Pothos’s (2007) review of the adult AGL literature, the distinction between explicit and implicit knowledge is a key consideration when testing adults. There are several measures that can address this question. First, if participants are consciously assessing the sequences then their confidence ratings ought to correlate with their accuracy. This was found to be the case with a highly significant point-biserial correlation, rₚ = 0.25, p<0.001 between participants confidence and their accuracy. This can be seen in Figure . Secondly, participants might be expected to respond faster on correct trials. This was indeed the case, participants took an
average of 2.9 seconds to respond when they were correct and 3.8 seconds when they were wrong. The reaction time data was positively skewed so these averages were calculated using geometric means. Likewise, the statistical significant analysis of this difference was performed using a one-way anova on the log-transformed reaction time data. The difference was found to be highly significant, \( F(1,334) = 9.70, p<0.002 \). One further point of interest was that there appeared to be no learning over the course of the experiment, overall accuracy rates were the same for the first six trials 64.3% and the last six trials 63.7%.

2.5.3. Discussion. Experiment 4 was designed as an approximate adult equivalent to Experiments 1 and 3, assessing how good adults would be at judging the relative randomness of sequences of looming coloured shapes, presented in the similar manner as those shown to the infants. The adults were also tested on several other comparisons not seen by infants, including a completely deterministic sequence of the six shapes. It was found that adults were better than chance when one of the sequences was completely random (i.e. 1 vs. 2, 1 vs 3, 1 vs 4) and also comparing pair-based sequence to the deterministic sequence. But were at...
chance with the other comparisons (i.e. 2 vs 3, 3 vs 4). These differences were not well accounted for by a model that compared the relative entropy of the different sequences. Nor were any sex differences found.

It is interesting that both infants and adults can discriminate random versus pairs (Experiment 1) but cannot discriminate pairs versus triplets (Experiment 3) as this suggests there is continuity between the tasks and between the age groups. Although the adult single alternative forced choice task is necessarily very different from the infant controlled looking time task, it is possible that in both cases participants are responding to the same statistical features of the sequences. In the infant experiments, there is strong evidence that local measures of redundancy and some awareness of the relative frequency of certain pairwise transitions are driving their performance. In the adult task, reports from participants who had completed the task indicated that many of them adopted a verbal strategy, (including notably the three most accurate participants.) Typically the participants reported that they repeated the names of the shapes or the colours which lead them to notice repeating patterns. Thus, at least some of the participants were relying on an explicit strategy that relied on working memory. It is likely other participants also used similarly explicit strategies and the reaction time and confidence data support this idea. (Despite the task instructions that participants ought not to think about their responses but "go with what seems right".) A future study might investigate this by including a dual-task that occupies working memory.

Across the 12 trials adults saw 24 unlabelled 30 item sequences with 4 different grammars all instantiated with the same shapes. It is, therefore, very difficult for the adults to learn anything globally about the grammars because of interference between the different grammars in the different sequences. This further supports the idea that adults use a strategy based on local cues. This would also be consistent with the results of Jamieson and Mewhort (2005) in which adult participants were required to accurately reproduce 8 item sequences of 6 coloured discs. They found that subjects actively exploited the local redundancy in the sequences and chunk-based features but not the global (grammatical) structure and that they showed no implicit (tacit) awareness of the grammars they were exposed to. If adults are explicitly exploiting local-redundancy and chunks then does the previously presented evidence that infants appear to also exploit these features suggest
that the infants’ processing is also explicit? What would this even mean? These questions are addressed in the general discussion below.

2.6. General Discussion

Four experiments were conducted that could be characterised as investigations of Artificial Grammar Learning without a learning phase. The first three experiments investigated five month old infants’ spontaneous preferences when presented with two visual sequential grammars of differing complexity whilst Experiment 4 required adults to judge which of two paired sequences was more random. The infant studies found strong evidence that 5-month-olds will look longer at more complex, less orderly grammatical sequences in certain cases. A detailed analysis of the actual sequences of items the infants saw and the exact point they disengaged from these supported the idea that they are not learning global properties of the grammars presented but respond to local redundancy and possibly also small fragments of the grammar. There was an unexpected main effect of gender with male infants looking longer overall in Experiment 1 and a trend in the same direction in Experiment 2 but this effect went away in Experiment 3. There are no clear theoretical accounts of such a difference and limited and inconsistent empirical support for longer looking or slow habituation in male infants. Therefore, no interpretation of this effect is offered. The adults in Experiment 4 showed the same failure and successes as the infants in Experiment 1 and 3 and were shown to be likely to be using explicit strategies based on local features of the test sequences.

In Experiment 1, five-month-olds spontaneously and consistently preferred a random sequence compared to a more predictable sequence (paired items) of the same shapes, in an experiment that matched the test phase of Kirkham et al. (2002). There was no pre-exposure and individual item frequencies were equal, so infants must look longer to the more random sequence because they are aware of some aspect of its greater complexity. Complexity analyses support this conclusion; infants looked away when the sequences exhibited greater zero and first order local redundancy; that is, when the sequences became more repetitive. Since this is more likely in a constrained grammar, this can account for why infants look away sooner on those trials. The relative frequency of items or pairs was not greater at the point of look away, suggesting that infants do not track overall item or pairwise
frequencies. Male infants discriminated better which may be because they looked longer overall.

Experiment 2 used sequences built from three pairs or two triplets of shapes to see if infants would look less at the less complex triplet-based sequences. Due to a procedural error, there was a very high degree of overlap between the two sequence types. In this case the infants showed a significant preference for the more complex sequence (pairs) only in the first block. There was also an interaction between order of presentation and the discrimination that was consistent with the idea that infants have spontaneous sensitivity to complexity. Whilst an asymmetry between the item-pairings in the triplet and pairs sequences could explain the interaction seen and may indicate that infants were learning features of the grammars presented. Additionally, there was a trend towards male infants looking longer.

Experiment 3 re-ran the task of Experiment 2 but with minimal overlap between the pairs and triplet based sequences. Infants looked equally long at both types of grammar. Adults in Experiment 4 also fail to discriminate these grammars, so it may be that the complexity differences between the two sequence types was below the threshold of the participants’ awareness. However, a secondary prediction of a global complexity was also not supported; when overall looking times were compared, participants did not attend less overall in Experiment 3 than in Experiment 1. When analysed in terms of whether local redundancy and the relative frequency of shapes and pairs of shapes could predict when infants would look away from the sequences, it was found that single item and pairwise redundancy measures were significantly above average at the point where infants looked away, particularly when considered over a six item window. As in Experiment 1, infants were more likely to look away when the most recent few items they had seen were more repetitive. The frequency analysis showed that infants looked away when they encountered relatively more frequent pairs of items. These analyses suggest that infants are sensitive to local redundancy and may be learning aspects of the grammar, responding to the more frequently encountered elements.

Experiment 4 was a control task, assessing how well adults could judge the relative randomness of sequences of looming coloured shapes, presented in the similar manner as those shown to the infants. The adults also saw a completely deterministic
sequence of the six shapes. It was found that adults were better than chance when one of the sequences was completely random and when comparing a pair-based sequence to a deterministic sequence, but were at chance with the other comparisons (i.e. pairs versus triplets, triplets versus deterministic). Whilst these two sequence comparisons had the smallest difference in relative entropy, overall accuracy was well accounted for by a binary logistic model that compared the relative entropy of the different sequences. Adult participants appeared to be using an explicit (often verbal) strategy that involved detecting local repetitions using working memory. This was consistent with earlier findings of Jamieson and Mewhort (2005).

2.6.1. Global or local? Statistics or chunks? Both or neither? Evidence from our studies suggests that global statistics do not predominate in either the infant or adult task. Certainly, there is no evidence of any rule learning in the sense of Pothos’s (2005) frequency-independent definition of rules. Rather infants appear to exploit local redundancy while adults use explicit strategies based on working memory. However, the results and analyses do not make it possible to distinguish if a chunk-learning may form a part of infant learning. Aslin et al. (1998) found that infants appear to compute conditional probabilities but as Perruchet and Pacton (2006) show this could equivalently be explained as chunk learning. It would be desirable to find some complexity analysis that considers the effect of chunk-like grouping of items.

In addition to introducing the calculation for local redundancy, Jamieson and Mewhort (2005) also introduced a measure which they called organizational redundancy, based Miller’s (1958) famous chunking hypothesis. This was intended to quantify the possible effects and the memory advantage of organizing information into chunks, (grouping the sequences into sub-blocks e.g. BDEGBDEA is easier to remember as BDE-G-BDE-A). They report evidence that participants do group the 8 item sequences in their task in this way. However, the measure they chose has a subjective factor because it depended on a post hoc selection of chunks that are considered, based on indications from participants of what chunks they claimed to have been using. This information cannot readily be obtained from the infants. For this reason, we did not consider using this formula in our complexity analyses. It may have been possible to perform the analysis using the chunks that we know
our structured grammars generate but that potentially begs the original question. Therefore, a question for future research might be to find an unbiased estimator for a chunk-based redundancy score and use it to see if this measure is predictive of infants’ disengagement from the sequence streams. Alternatively, Perruchet and Peereman’s (2004) PARSER claims to model the random formation of chunks and their consolidation or forgetting based on an associative memory mechanism, potentially this could be adapted to fit infant performance.

2.6.2. Implicit or explicit processing? There was strong evidence from the confidence data, reaction times and verbal reports that adults in Experiment 4 process the sequences explicitly. Whilst the local redundancy measures indicate that infants look away when they encounter repetition, suggesting that their performance is driven by short term memory effects. As Pothos’s (2007) review makes clear, the contrast between implicit and explicit processing is likely to be crucial to reconciling conflicting findings in the adult AGL literature. It is plausible that this dichotomy has an analogous importance in the interpretation of infant AGL results. A parallel might be found with the work of Rovee-Collier (1997) within infant long term memory research which rejects the traditional view was that infants did not possess explicit, episodic memory, showing instead that this develops in parallel with implicit systems in the first year life. Less is known about the capacity and development of short term and working memory in infancy but this may have bearing on the current studies.

Diamond (1985; 1990) found for infants under 7.5 months any delays disrupted their reaching performance on the A not B task, whilst older infants can gradually accommodate greater delays up to about 10 seconds at one year. Whereas, Gilmore and Johnson (1995) used an oculomotor delay response task and found that 6 month olds could retain information about stimulus location for about 3 to 5 seconds. Haith and colleagues (Haith, 1993; Haith et al., 1996; Adler and Haith, 2003) have found that infants as young as three months will learn to anticipate regularities in spatiotemporal sequences, making anticipatory eye movements when a sequence of stimuli alternated on left and right of a screen. The infants can succeed in integrating information over several seconds and anticipate better when some properties of one of the stimuli are held invariant and others vary, such as picture
content (Wentworth and Haith, 1992) or item (Adler and Haith, 2003). Those experiments prove that even young infants can engage with sequential stimuli and form expectations about repetitions in the sequences. Which in turns gives support to the idea that infants in Experiment 1, 2 and 3 were using explicit cognitive processes.

These results point to the likelihood that there are multiple processes involved in infant artificial grammar learning, not just one statistical learning mechanism but several. This suggests that it is methodologically wrong to design experiments that attempt to control all sources of ‘extraneous’ variation to give a highly simplified and constrained task that might not capture the range of infant ability. We suggest that a better approach is to retain the complexity in the stimuli and to collect rich datasets and then attempt to constrain and control for variation in the analysis phase by conducting multiple sophisticated analyses on the same experiment. A further step in this direction is to use explicit models such as neural network models that fit actual real-world data. Previous unpublished work by this author (Addyman, 2005) found that if anything neural networks were too good at learning the Kirkham et al. (2002) task. Even simple perceptrons rapidly detect the global statistical regularities but could not account for the local effects seen in these studies. In this respect, a more biologically inspired model may fit better the actual performance seen. In this respect, the Habituation, Autoassociation and Brain (HAB) model of Sirois and Mareschal (2004) is a good candidate. This is a connectionist model of habituation based on an explicit implementation of an architecture based on the relevant brain areas (hippocampus, thalamus and cortex). See Káldy and Sigala (2004) for a review of brain areas likely to be involved in visual working memory in infancy.

One of the main conclusions of the Kirkham et al. (2002) study was that statistical learning was a domain general process. These experiments support the idea that as infants can discriminate sequences in the visual domain, their performance can be explained by response to relatively localized simple features rather than sophisticated grammatical learning. It is possible that these may be processes specific to the visual domain. Whilst the processing of linear sequential information is the hallmark of the auditory system, rarely are stimuli of this kind found in the visual domain. Therefore, care must be taken about making the inference back from
these results to experiments in the auditory domain and future work would need
to investigate the effects found here directly with appropriate auditory stimuli.

This chapter has examined in considerable detail a set of relatively simple artificial
grammar tasks and found evidence that both infants and adults may discriminate
these stimuli in similar ways. The results point towards a primarily associative pro-
cess heavily dependent on working memory. As was seen in the survey of theoretical
perspectives on artificial grammar learning (section 2.1.2), Pothos (2007) empha-
sises that two dimensions are pertinent to the study of AGL, the implicit-explicit
distinction and extent to which the tasks and theories address questions about rules
and generalisability. Whilst these results have shown that the implicit-explicit di-
mension is of direct relevance to the study of infant AGL, it has ignored the rules
dimension. Gómez and Gerken (2000) also emphasise the importance of questions
about pattern- and category-based abstraction. The next chapter investigates ab-
straction and generalisation, albeit not in an AGL context. Rather the focus is on
infants’ appreciation of abstract Same/Different relations in two novel paradigms.
However, the recognition of the general property of identity or difference in the
two element pairs AA or AB is a prerequisite for infants’ success at generalising
in AGL tasks such as Marcus et al. (1999). Therefore, the approach to infant rule
learning presented in the next chapter is directly relevant to and in continuity with
this existing research into infant Artificial Grammar Learning.
CHAPTER 3

Infant learning of abstract Same/Different relations.

3.1. Introduction

One fundamental characteristic of human cognition is an ability to make use of abstract concepts and perceive similar relations between otherwise unrelated items. These can take a very wide variety of forms including judgements of numerosity and of comparative quantity and size (largest, smallest, etc), relations expressed by spatial prepositions (above, behind, between, etc.) and numerous others. See, for example the whole of the first section of Roget’s Thesaurus which is devoted to ‘Words Expressing Abstract Relations’. One of the simplest examples is the two item same/different (S/D) distinction. If we take any set of 9 distinct types of items labelled A to I and consider the pairs of items AA, BB, CC, DE, FG, HI, the first three pairs share the relationship of sameness and (by contrast) the last three have the common property of difference. This relationship holds at an abstract level, irrespective of the actual items used and it is this abstraction that defines the relationship. Despite this apparent simplicity, it has been difficult to demonstrate with species other than higher mammals. For example, it has taken extensive experimentation to demonstrate unequivocally that pigeons are capable of solving this problem (Blaisdell and Cook, 2005). Since the failures of certain species on seemingly simple tasks can be as instructive as successes it is useful to consider the comparative literature in some detail. Therefore, this chapter begins with a review of the debate in the comparative literature before surveying the much more limited infant work and introducing two new experimental investigations with human infants.

3.1.1. Animal studies. For a long time, animal researchers had difficulty demonstrating S/D discrimination in pigeons or rats. Zentall and Hogan and col-
3.1. INTRODUCTION

leagues (Zentall and Hogan, 1974, 1976; Zentall et al., 1981) demonstrated that pigeons could use some aspect of stimulus similarity or difference in a match-to-sample task. For example, in Zentall and Hogan (1974) birds were trained on the colour match or mismatch between cue and target stimuli (e.g. a red or green cuing lamp matched/mismatched to red or green target lamps) and tested for their ability to transfer this to new stimuli (e.g. blue and yellow lamps). Speeded transfer was found when second task was congruent to the first (e.g. both matching or both mismatching). In a review article, Premack (1983) pointed out limitations in the match-to-sample technique. These techniques do not necessarily provide a good test of the animals ability to use the abstract relational properties of the stimuli because the animal may simply be responding to the item it has seen before. This is a different question to one of identity and a confound that cannot easily be controlled. Meanwhile, Iversen (1997) found that the spatial location of a cuing stimulus had a strong effect on rats ability to solve even these match-to-sample tasks.

Premack (1983) proposed that generalization in the two (or more) item S/D task is better suited to testing genuinely abstract S/D concept learning. This has proved difficult with animals. Pearce (1988) found that pigeons could classify patterns based on prototypes determined by the heights of three different coloured bars but that they failed to learn the task when relation was of equality or difference in height between just two bars. With a modified training regime his pigeons were able to perform the match (Pearce, 1991). However Pearce believed that this could have been a consequence of them being able to memorizing all the possible configurations. This conflict between learning item-specific cases and acquiring a generalized response is a common problem in pigeon categorization experiments and there will always be some trade off between the two (see Wright, 1997 for an analysis of this type of conflict). Therefore the crucial measure of performance will be the response to completely novel stimuli. However, with many designs this is not possible.

Blaisdell and Cook (2005) reported a successful demonstration of two item S/D learning in pigeons where the learning stimuli were coloured geometric shapes and the pigeons were tested on transfer to novel coloured shapes. The training stimuli were taken from a set of six shapes, each in one of six colours. On a single training
3.1. INTRODUCTION

trial a same pair (e.g. two red squares) and a different pair (e.g. a purple chevron and a green star) were displayed on either side of computer monitor and half the pigeons were reinforced to peck the same pair and half to peck different pair. Six out of 6 pigeons reached an 80% criterion within 28 training sessions. Pigeons were then tested on transfer to either novel shapes, novel colours or both novel. Transfer occurred in all conditions but was weakest in the novel colour (i.e. shape-same) condition and strongest when both cues were changed. This provides strong evidence that pigeons can transfer in both dimensions, though performance was driven primarily by colour and the dimensions seemed to be treated independently. Thus it seems that although pigeons can make some abstraction based on the S/D relation this seems to be tied to individual dimensions rather than taking place at a more 'conceptual' level.

Further evidence in support of a relatively low-level perceptual account of pigeons abilities comes from Young and Wasserman (2002). In their experiment, pigeons responded to a 16 item array of computer icons that varied in entropy from low (all same) to high (all different), responding was measured in a go/no-go task where arrays with higher or lower entropy were selectively reinforced. Pigeons trained to respond positively to high entropy arrays showed better discrimination. Young and Wasserman provided analysis which supported the interpretation that the pigeons multi-item S/D judgements followed a logarithmic (rather than linear) entropy function, suggesting that same-difference could be considered a dimensional rather than a categorical distinction in certain cases.

Recently, Wright and Katz (2006) ran a comparative study with pigeons, rhesus monkeys and Capuchin monkeys using sets of photographic stimuli which provided strong evidence of learning and generalization in all species. Rhesus and Capuchin monkeys learned at comparable rates while pigeons learned more slowly. The animals were trained on vertically arranged pairs of pictures selected randomly from a training set. If the two pictures were identical the animal had to tap on the lower picture, if they were different the animal had to tap on a target square to the left of the lower picture. Correct responses resulted in a food reward. All species reached a criterion of 80% with a training set of just 8 items (giving 8 same item pairs and 56 different item pairs). However, no animals initially learned to transfer this to novel test items. Increasing the size of the training in stages lead
to full generalization (when accuracy with novel items reached equivalent accuracy as baseline) at 128 items for the monkey species and 256 items for pigeons. Thus there were quantitative but no qualitative differences between species and a strong criterion for abstractness (equivalent performance with novel material) was met. Wright and Katz speculated (based on human similarity ratings) that this continuity would extend to higher mammals. This appears to be supported by Mercado et al.'s (2000) demonstration of S/D learning in two bottlenose dolphins, albeit using a very different paradigm.

Likewise, in a widely cited study, Oden et al. (1990) tested infant 11 month old chimpanzees with several two item S/D tests. In their first experiment Oden et al. found that the infant chimpanzees would look longer at a novel over a familiar item, indicating that they do at least perceive the similarity between the objects. Their second experiment tested the perception of Sameness and Difference relation. The chimpanzees were given five minute sessions with pairs of toys to physically examine. The two toys were attached to a board and could either be the same or different from each other (e.g. AA or CD). Immediately afterwards the chimpanzees were given a new pair of toys that maintained or changed the relationship. In four separate trials each chimpanzee saw all four possible transfer relations (e.g. AA then BB, AA then EF, CD then BB, CD then EF). There was a main effect of familiarity, so that looking decreased more when the second pair had the equivalent relation as the first pair. However, in the third experiment the chimpanzees failed on a matching to sample task where they were required to physically match a pair of items to another pair demonstrating a similar relationship. In initial training, the chimpanzees succeeded in reaching a 10/12 criterion but they could not transfer this to novel test sets. This contrasts with the monkeys and pigeons in Wright and Katz (2006). Oden et al. speculate that it was the instrumental aspect of the task that led to the failure with the infant chimpanzees unable to integrate perceptual awareness with motor planning.

Many of these comparative results are of interest from a developmental perspective when trying to predict how human infants would perform. On one hand, the success of a wide range of species at generalising the S/D concept suggests that human infants ought to be able to both perceive and act on the abstract relational information and that, at least in the case of perception, there is no reason to
suspect that there is any strong prediction about how early in development this competence could be found. Furthermore, infants might be expected to perceive the S/D relation when tested with either rich photographic (Wright and Katz, 2006) or impoverished geometric stimuli (Blaisdell and Cook, 2005). On the other hand, Young and Wasserman’s (2002) study shows that care must be taken in giving too rich a conceptual interpretation to S/D discrimination. Blaisdell and Cook (2005) are careful to emphasise the importance of generalization as a measure of a genuinely abstract conceptual. While the failure of Experiment 3 of Oden et al. (1990) emphasizes the difference between perceptual awareness and instrumental knowledge and in particular highlights how lack of integration between developing systems may lead to failures on certain tasks. Nonetheless, the overall message from the comparative literature is that one might expect infants to succeed on a range of S/D tasks and demonstrate genuine abstract S/D concept providing care was taken with the design and demands of the task. The next section looks at previous relevant studies with children and infants.

3.1.2. Infant and child studies. Only two studies (Tyrrell et al., 1991, 1993) have directly investigated the S/D concept with human infants, although a concept is implicit in a number of other contexts, most notably in artificial grammar tasks like Marcus et al. (1999). An earlier study (Caron and Caron, 1981) found that infants habituated to matching pairs of items dishabituated to non-matching pairs. However the study was not designed specifically to look at S/D concept learning and because infants were only ever habituated to sameness, non-conceptual explanations could not be ruled out, such as the infants responding to the greater complexity or the broken symmetry of the novel Different display.

Tyrrell et al. (1991) designed their study to address this problem. Using a similar design to Experiment 2 of Oden et al. (1990), Tyrrell et al. tested 29 week old infants using a preferential looking paradigm. Each infant took part in two familiarization/novelty preference test, one a concrete test and one testing the abstract relation. In the concrete test, the infants were familiarized with pairs of small toys affixed to a board, infants in the Same condition had two 20s exposures to a pair of matching toys (AA) and infants in the Different condition saw two dissimilar toys (BC). All infants then saw pairs AA and BC next to each other and in both groups
they preferred the unfamiliar pair. In the second abstract phase of the experiment, the infants in the two groups were familiarized to a new pair of items with appropriate relation (DD or EF respectively) and tested for preference for new items GG vs HI. Infants in both groups looked significantly longer at the pair that had a novel relation relative to familiarization. There were no interactions between conditions, although there was a main effect that infants in the Same condition looked longer overall. The fact that infants in both conditions responded to the novel relation in the abstract test suggests that they are responding at a conceptual level. However, the very limited familiarization phase showing just one pair of items in each case and using ‘toys’ constructed out of a common set of parts increase the likelihood that the response is to some concrete similarity across conditions than through the use of a completely abstract S/D concept.

Tyrrell et al. (1993) used a conditioned head-turn paradigm with a pair of matching toys on one side and a pair of mismatching toys on the other. Turning to the correct side caused a recording of a children’s story to start playing. Infants were assigned to one of four conditions and were reinforced over three separate blocks of four, four and eight trials each, with an new set of objects on each block. Infants in one condition were always reinforced for looking at the same pair, whilst in another condition were always reinforced for looking at the different pair. Infants in the two other conditions were reinforced in the same way for blocks 1 and 2 but were then switched for the final block to be reinforced for the opposite relation. The results supported the hypothesis that the infants were learning the concept rather than simple association to particular items; All groups increased their looking to the correct side over the first 2 blocks but the infants in the switch conditions were impaired at the start of the 3rd block. Tyrrell et al. (1993) argued that this pattern of results mirrors the reversal shifts found in discrimination learning and fits the predictions of the House et al. (1974) model of discrimination learning. They also consider the conditioned head turn paradigm to be an instance of instrumental learning and so draw a contrast between the success of human infants in their task and the failure of the infant chimpanzees in Oden et al. (1990), although they acknowledge that the different task demands make it impossible to make a direct comparison.
3.1. INTRODUCTION

An awareness of S/D contrasts is a necessary pre-requisite for success in the artificial grammar tasks of Marcus et al. (1999). In their Experiments 1 and 2 it is also a sufficient condition. Seven and a half month old infants were habituated to an artificial grammar of syllables of synthesized speech. Infants who had been exposed to several minutes of an ABB tri-syllabic pattern showed preference for ABA patterns over ABB patterns when tested with completely novel syllables. An awareness of the presence of a reduplicated syllable would be sufficient to discriminate between the two types of grammar. In Marcus et al. (1999) Experiment 3, the contrast was between an AAB and ABB pattern and a simple S/D discrimination would no longer be sufficient. Equally informative are studies such as Marcus et al. (2004) where the AAB/ABA/ABB finding was not replicated when the stimuli were pure tones. This was interpreted, by the authors, as support for a domain-specific linguistic ability. However, Saffran et al. (2007) after replicating the Marcus et al. (1999) findings using pictures of dogs and cats, speculate that the stimuli must themselves form familiar categories before infants are able to map relations from one case to another. It is an open question as to the extent to which S/D discrimination is a more general process or relies on pre-existing categorical knowledge.

Also of interest in this respect are the result of Kotovsky and Gentner (1996), who investigated the abilities of 4, 6 and 8 year old children to recognize relational similarities. Children were presented with triads of objects that demonstrated some relation and where then tested on a two alternative forced choice task to find the equivalent relation. The test pairs could vary on dimension and on polarity. For example, if ”xXx” were the initial configuration, participants would be given the choice of ”oOo” or ”ooO” in the unchanged case, a choice between ”OoO” and ”oOO” in the changed polarity case and a choice between ”121” and ”112” in the changed dimension case. (In the actual experiment, the stimuli were triads of circles and squares which varied the dimensions were size and colour and the polarities of large and small and light and dark.) Six and 8 year olds were responding above chance in all conditions but 4 year olds were only capable of answering correctly when dimension and polarity were unchanged. However, providing them with coaching that drew attention to these dimensions without explicitly highlighting the commonalities improved their performance on the generalization task. Kotovsky and Gentner (1996) called this the progressive alignment hypothesis suggesting that
experience with concrete similarity leads to success with more abstract material because an alignment of the structures makes the relation more salient. Gentner and Medina (1998) develop this position further claiming that structural alignment is the process by which rules are acquired.

Given this review of the literature, the present study has several aims. The primary objective to test pre-linguistic infants using new experimental designs that will be suitable for demonstrating that these infants can make use of abstract S/D concept to discriminate certain stimuli. In particular, it is hoped the infants will meet the criteria of Blaisdell and Cook (2005) of equal performance with unfamiliar materials. An additional aim is to be able to demonstrate some continuity with the animal literature using tasks that have some common elements (allowing for necessary differences in experimental design required for testing such different populations.) Thus the experiments will use both rich photographic stimuli and simplified coloured, geometric shapes. The design will include a perceptual and an instrumental task to explore the differences that were observed in Oden et al. (1990) but which were not found in the work of Tyrell and colleagues. Additionally, the study aims to address the questions raised by Saffran et al. (2007) about whether pre-existing category knowledge and/or domain specificity are important for S/D learning. Another aim of the present study will be to test the same infants using several different measures. Success on S/D tasks with very different paradigms will provide converging support for a conceptual interpretation of the behaviour. With sufficient data it might be possible to compare individual differences in performance from one task to another, if there no correlation between success in one task and another, the case for an abstract conceptual mechanism is weakened. Finally, as ought to be the case with any infancy study, we are interested in the development of these abilities. How early do competencies arise and how do they change over time?

Taking all these factors into consideration, two tasks are presented that look at S/D learning. One task will be a simple habituation/dishabituation paradigm using photographic stimuli. The items will be unrelated to each other, not forming an particular category and infants will be habituated to either a same or a different condition as a between subjects variable. Another task will adapt McMurray and Aslin’s (2004) anticipatory eye movement paradigm to assess infants’ learning of
associations between Sameness leading to one outcome (e.g. an occluded pair of objects reappears on the left) and Difference leading to another outcome (e.g. objects reappear on the right). Here, the stimuli will be simple geometric shapes and both learning and generalization will be assessed. The requirement for the infant to actively anticipate the correct side makes this a task that investigates instrumental learning. The same infants will be tested on both tasks in order to get data for a comparative or individual differences analysis. For example, correlations between performance on both tasks could be suggestive of a single underlying process. It is hypothesized that 8 month old infants will succeed at both tasks and propose to follow up with a different age group depending upon the results of these first two tasks.

There is a methodological challenge when testing young infants that they cannot concentrate for extended periods. This is a danger when trying to collect data from multiple tasks. Although one aim of the present study was to get all infants to complete both experiments in one visit in the hope of obtaining data for between task comparisons, the main priority was to obtain good data within each task. Furthermore, practical constraints meant that infants could not be invited back. Therefore, the experiments were completely swiftly with the infant in the testing area for the shortest time possible. The infants were given a short break between the two experiments, returning to the reception area. The studies were in different testing rooms with different layouts to reduce the likelihood of infant adapting to the testing environment.

3.2. Experiment 1a - S/D Discrimination in 8-month-olds using Photographic Stimuli in a Habituation Task

Experiment 1a adapts a standard habituation/dishabituation paradigm to test infants’ sensitivity the abstract Same/Different relation. Eight-month-olds are chosen because the success of 7.5-month-olds in Tyrrell et al. (1991) and Marcus et al. (1999) suggests that infants of this age should succeed on a direct test of S/D discrimination. A set of unrelated photographic stimuli is chosen to give a comparison to the work of Wright and Katz (2006) whilst being able to address Saffran et al.’s (2007) hypothesis that infants will not succeed on such tasks unless the objects form a familiar or coherent category for them.
3.2. EXPERIMENT 1A - S/D DISCRIMINATION IN 8-MONTH-OLDS

3.2.1. Method.

3.2.1.1. Participants. A total of 15 eight month infants provided data (M= 249 days, range 232-259 days, 6 female, 9 male ). All were full-term infants with no delivery complications. A further four infants were tested but were excluded due to fussiness (2) or experimental error (2). Infants were recruited via the CBCD’s participant database. Parents were not paid for their infants’ participation but infants were given a certificate, a small present and travel expenses were reimbursed. Parents were informed of the nature of the study and their written consent was obtained prior to testing. Infants took part in both Experiment 1a and Experiment 1b in a single visit. For each infant it was randomly determined which experiment he or she would take participate in first.

3.2.1.2. Apparatus. A Macintosh G4 computer running MATLAB R2006a (version. 7.2) with the PsychToolBox (Brainard, 1997; Pelli, 1997) was used to control the display of the stimuli and monitor the habituation criterion, using experiment control routines written by the author. The stimuli were displayed on a Samsung 42” plasma screen and infants reactions were filmed and recorded using an infrared DV camera for later offline coding.

3.2.1.3. Stimuli. The stimuli were drawn from a set of 25 coloured photographs of inanimate objects, as shown in Figure 3.1. All the pictures were scaled to be approximately the same size (300x300 pixels). Pairs of items were displayed on a uniform mid-grey background at a screen resolution of 768 x 1040 pixels with their centres 520 pixels apart, which equated to a visual angle of approx 10° per item at a separation of approx 32°.

3.2.1.4. Procedure. At the start of the experiment the infants and their caregiver were invited into the testing room and were seated at a distance of approximately 1.3m in front of the screen with the middle of the screen at approximately eye-level for the infant. The infant was seated on the care-giver’s lap and the caregiver was instructed not to interact with their infant during the experiment. The experimenter was seated at a control desk in the same room as the infant but hidden behind a heavy curtain. The lighting was kept low and quiet classical music was played in the background throughout the experiment so that the testing room
was not completely silent. This was because it had been observed in previous studies that infants sometimes become agitated when moved from the bright and busy reception area to a dark and silent testing booth.

When the infant was settled, the habituation phase began. The screen started as a blank grey slide and each trial began with a 'boing' sound effect. Each slide then appeared on the screen and remained there whilst the infant continued to look at it or until a maximum of 20 seconds had elapsed. An infant was judged to have looked away if either they looked away for a single period of 500ms or if they accumulated 1000ms of glances away from the screen during a single trial. Looking behaviour was coded on-line by an experimenter holding a key down on the computer whenever the infant was watching the screen, with the control software determining when a look away criterion or a maximum time was reached. There were a maximum of 19
habituation trials. Habituation was deemed to have occurred if the average length of last two trials was less than 50% of the average of the first two trials.

The stimuli displayed at habituation and test depended on the experimental condition. Infants were randomly assigned to one of two conditions; half the infants were in the *Same* condition, half were in the *Different* condition. For infants in both conditions, each of the habituation trials consisted of a slide featuring pairs of items, the order and selection of items determined by a pseudo-random permutation routine. The experiment employed a 'yoked' design, so that the same sets of items were used in habituation and test for one member of each group. A random subset of 19 items where used in the habituation trials. For infants in the *Same* condition, each stimulus consisted of an item from this set paired with itself. For infants in the *Different* condition, the item on the left was always paired with different item on the right. This was achieved by using an identical order of items for both sides but shifting one sequence by 7 places. So that the item that appeared left position on the n-th trial would re-appear on the right position on trial n+7 (modulo 19). This ensures that the infants in both conditions would see the images with equal frequency overall. For example, an infant in the *Same* condition might see two pictures of a watering-can simultaneously whilst the yolked infant in the *Different* condition would see the watering-can in two separate habituation trials. For each infant, the remaining 6 pictures were used to produce the test items, two *Same* test pairs and two *Different* test pairs. The infant in the *Same* condition saw the *Same* test items first then the *Different* test items, the yoked infant the *Different* condition saw the identical test images but in the opposite order. Figure 3.2 gives a schematic representation of the procedure.

### 3.2.2. Results.

A total of 15 infants successfully completed the study. Looking time scores were calculated by off-line coding of the video recording of each experimental session, account was taken of the time that the total time the spent infant looking at the screen, with glances away during a given trial deducted from the totals. A second experimenter, blind to the experimental hypothesis, double coded the videos of a randomly selected 20% of the infants (three infants). A Pearson correlation on the two sets of data gave a high degree of inter-rater reliability,
Figure 3.2: Procedure for Experiment 1. Infants were randomly assigned to either the Same or Different condition. After being habituated to pairs of items demonstrating the relevant relation, they were shown four test trials. The first two had the same relation as the habituation trials, whilst the last two had the opposite relation.

$r=0.96, p<0.001$. For each infant the average of their looking times for first two habituation trials (Av. Habit 1 & 2) was taken and compared it to the average for the last two trials with the same relation as in habituation (Av. Habit Test) and for the two novel relation trials (Av. novel test). The results are summarised in Figure 3.3. Seven of the infants completed Experiment 1a first, whilst the remaining 8 took part in Experiment 1b first. It is possible that fatigue may have had an influence on performance so a preliminary mixed 3x2x2 ANOVA on looking times was performed with Trial Type as a within subject factor at three levels (Av. Habit 1 & 2, Av. Habit Test, Av. novel test) and experiment condition and experiment order as a between subjects variables each at two levels (habituated to Same pairs or Different pairs and Experiment 1a or Experiment 1b run first). This analysis showed that the order variable had no main effect $F(1,13) = 1.1, p > 0.31$ or interaction with other variables, $F(2,26) < 1$ and so it was excluded from further analysis. A similar analysis showed a marginal effect of gender, $F(1,13) = 2.93, p< 0.11$ and an almost significant trial x gender interaction $F(2,26) = 2.90, p < 0.073$. Although this may be attributable to uneven cell sizes, this variable was retained in subsequent
The experiment was concerned with whether the infants significantly increase their looking to the test trials showing the novel abstract relation, therefore a mixed 3x2x2 ANOVA was conducted with trial type (Habit Trial 1 and 2, Habit Test, Novel Test) as the within subject variable and Experimental Condition (Same or Different habituation trials) and render as the between subject variables. In line with the hypothesis there was a highly significant main effect of trial type, indicating strong dishabituation to the novel test trials, $F(2,22) = 9.8$, $p<0.001$. There were no other significant main effects. Crucially, there were no significant interactions indicating that the result was not driven by infants in the Same condition responding to the greater complexity of test displays with two Different items. Similarly, there was no main effect of condition, $F(1,11) < 1$, indicating that overall looking times were comparable between these two conditions. The marginal trial x gender interaction remained $F(2,22) = 2.26$, $p < 0.093$, but was not investigated further because there was no clear interpretation and it may be due to the small sample sizes.
To confirm that the infants were dishabituating to the novel relation in both conditions, planned comparisons were performed (related unidirectional t-tests). These indicated that the infants who had been habituated to Same pairs dishabituated to the Different test items, $t(7) = 2.78$, $p < 0.02$. The infants habituated to Different pairs looked longer at test pairs with two Same items, $t(6) = 2.08$, $p < 0.04$.

3.2.3. Discussion. The results of this experiment provides strong evidence that 8 month old infants are sensitive to the abstract relation of Same/Different. Infants who had been habituated to pairs of same or different objects dishabituated when shown pairs of new objects with a novel relation. Furthermore, there were no differences between the two conditions, infants who habituated to dissimilar pairs were just as likely to dishabituate to the novel relation. This supports previous findings that 7.5 month olds succeed on artificial grammar tasks (Marcus et al., 1999; Saffran et al., 2007) and that 7 month olds can perform some simple S/D discriminations (Tyrrell et al., 1991, 1993). The study also provides continuity with the findings of Wright and Katz (2006) showing infants are also able to make a S/D discrimination using photographic stimuli.

This continuity with the animal results and the strength of the effect in infants suggests the mechanism is more general than previously believed. Saffran et al. (2007) speculate that an awareness of category membership was a pre-requisite for noticing higher level commonalities. But in this case, although all the items were photographs of inanimate objects, there was no more specific perceptual or conceptual category that they belonged to. Yet the infants were clearly sensitive to the S/D relation despite the lack of common category for all the presented items. Furthermore, this experiment provides stronger evidence than the experiments of Tyrrell and colleagues where in each case there were only a very small set of items, making it unclear that a genuinely abstract relation was being detected.

It was perhaps surprising that there was no differences in overall looking time between the conditions. It might be expected that the Different slides are intrinsically more interesting than the redundant Same slides and that therefore infants in the Different habituation condition would look longer, especially at the start of the experiment. One possible explanation as to why the greater complexity did not lead to longer looking might be that any such effect was offset by the greater saliency
of matching pairs of items. Thus in the former case infants have more individual features to look at but in the latter they are more inclined to make comparisons between the objects. Whatever the explanation, the fact that both infants familiarised with pairs of Same items and infants familiarised with pairs of Different items looked for similar amounts of time, habituated in both cases and dishabituated equally to pairs of items showing the opposite relation. This is strong evidence that these infants were responding on an abstract level to the S/D relation.

3.3. Experiment 1b - S/D Learning and Generalization in 8-month-olds using an Anticipatory Looking Paradigm

This experiment adapts the anticipatory eye movement paradigm developed by McMurray and Aslin (2004). In particular building on the method of their Experiment 3. The eye-tracker can record infants eye movements and detected which part of the screen they are looking at a given point in time. Therefore, it is possible to use this technology to see if the infants will learn to anticipate the outcome of a repeated event. By tying spatially distinct events to different cuing stimuli it is possible to see if the infants get faster at looking towards the correct cued location. This would indicates that infants can discriminate the cuing stimuli. A generalization or conflict phase can then be introduced with new stimuli that have some or all of the properties of the learning phase stimuli. How the infants respond to these indicates if they are able to generalize or (in the case of a conflict trial) which dimensions dominate in classification. For example, in Experiment 3 of McMurray and Aslin (2004), 6 and 7 month old infants were trained with red squares moving to the left and yellow crosses moving to the right behind a light blue inverted-T occluder. They were then tested to see if they had learnt this relationship and how they responded to conflict trials featuring a yellow square or a red cross. It was found that the infants who learned the initial relationship generalized on the basis of colour over shape. That is, they anticipated the yellow square would emerge on the right and vice verse.

The current experiment makes several methodological changes to the McMurray and Aslin paradigm. One of the biggest changes was to reduce the speed with which the stimuli moved across the screen. In the original experiment, stimuli were fully occluded for only 750ms and each trial only lasted a maximum 6000ms. One
problem here was that infants were equally likely to expect the stimuli to appear along a linear trajectory directly above the occluder as they were to have looked either left or right, consistent with the actual paths the stimuli took. It is possible infants have a pre-potent expectation of linear motion (see von Hofsten et al., 2000).

Increasing the length of the occlusion allows time for the infants to overcome any such bias and search in the correct location. Therefore, in the current experiment the stimuli move slower. In particular, they remain occluded for 3000ms, with the whole animated sequence lasting approximately 6000ms. An additional possible benefit is that slower moving stimuli that remain on the screen for longer allow more opportunity for the infants to encode the two items that are presented and their relationship. A further problem of the original study was that infants missed some trials because they were not looking in the right direction. For the infant to learn the dependencies, it is essential that they are attending to the screen. Therefore, in the present experiment, the software only started each trial when it determined that the infant was looking at the appropriate lower middle third of the screen. Finally, the number of learning trials is increased to give infants the fullest opportunity to learn the relation. The full procedure is explained below.

3.3.1. Method.

3.3.1.1. Participants. The same infants who took part in Experiment 1a also took part in Experiment 1b. A total of 17 infants (8 female) were tested, however due to difficulties with calibration, the sensitivity of the eye-tracker and the length of the experiment, not all data was usable. Four infants provided almost no looking time data, 3 infants provided only a small amount of data that was usable in group analyses but not individually. Of the remaining infants, 6 provided good data whilst for the remaining 3 timing and synchronization problems meant that it was not possible to be sure of exactly when the relevant occlusion events began and ended, these infant were therefore excluded.

3.3.1.2. Apparatus. The experiment used a Tobii 1750 eye-tracking camera with integrated 17" LCD screen. It was connected to a Dell PC running Windows XP. The presentation of stimuli was controlled using Exbuilder software originally developed in Dick Aslin’s lab at the University of Rochester, NY and modified for this experiment by the author and Fani Deligianni from the Centre for Brain and
3.3. EXPERIMENT 1B - S/D LEARNING IN 8-MONTH-OLDS

Cognitive Development, Birkbeck, University of London. A digital video camera placed directly above the screen recorded the infants reactions.

3.3.1.3. Stimuli. All the infants saw the same stimuli. An inverted T-shaped red occluder on a black background was present on the screen throughout each block. The shape and location of which can be seen in Figure 3.4. On the screen it was approximately 23.3cm across and 11.6cm high at its maximum, subtending visual angle of $21.9^\circ$ by $11^\circ$ when viewed at 60cm. For the learning phase the paired stimuli were a yellow circle and a light blue square. For generalization phase shapes included a maroon cross, an orange heart, a pink diamond and a green triangle. Each was about 2.5cm across on the screen, making a visual angle of approximately $2.4^\circ$.

3.3.1.4. Procedure. Each infant was tested individually and depending on a random assignment, either completed Experiment 1a or 1b first. Before the experiment began, the purpose of the study was explained to the caregiver. An explanation of how the eye-tracker worked was provided and it was emphasised to the caregiver that the infant ought to be kept still in one position throughout the experiment. At the start of the experiment, the infant and his or her caregiver were brought into the testing room and seated in front of the Tobii 1750. A preliminary phase of the experiment required the eye-tracker to be calibrated for each of the individual infants. A series of looming blue circles appeared at various locations on the screen to allow calibration of infant looking. Due to inattention and babies moving around, it was sometimes necessary to repeat this step to get accurate calibrations for all areas of the screen. After calibration the main phase of the experiment began.

The experiment was divided into four blocks, each of sixteen trials. The first three blocks were learning trials and featured only two of the stimulus shapes (the yellow circle and the light blue square) in all four possible stimulus pairings; two yellow circles or two blue squares, a circle and a square, or a square and a circle. Each pairing occurred four times in each block of sixteen learning trials with a different randomized ordering for each infant and each block. In the generalization phase, eight novel pairings were introduced four of which featured all pairs of maroon cross and orange heart and four of which combined the pink diamond and the green triangle. These were randomly mixed with eight trials that featured the
circle and square from the learning phase. Throughout the experiment, whenever the two objects were the same as each other they would always move to one side (determined randomly for a given infant) and if they were different they would move to the other. The experiment sought to determine if the infants would learn this dependency and if they could generalize it for the novel shapes in the final phase.

The experimenter attempted to get the infant to complete all 64 trials but this was rarely possible. Several steps were taken to increase the attentiveness of the infants. If necessary, some randomly moving cartoon figures were occasionally displayed to reorient infants to the screen. Furthermore, in the third learning block the parents were instructed to talk to their children, making a ‘peek-a-boo’ style game out of the occlusion sequence. They were encouraged to use phrases like “Where is it?”,”“There it is!” that would maintain the infant’s engagement but without giving the infant any information about the trial outcomes. Caregivers were silent at all other times. Finally, if the experimenter felt the infant was still becoming fussy or otherwise disinterested during the learning phase, the experiment was moved directly into the generalization phase. (Pilot testing found that the introduction of the new shapes in the generalization phase helped re-engage infants interest.) However, the experiment was terminated if the parent or the experimenter felt the infant was no longer capable of attending.

The procedure for a given trial is illustrated in Figure 3.4. A bell sound effect marked the beginning of each trial when the two objects first appeared at the lower middle part of the screen. If the eye-tracker detected the infant was attending to that part of the screen the objects would begin to move. If the infant was not attending to that part of the screen, the objects remained in place and loomed smaller and larger at a rate of about one full cycle per second until they captured the infants attention. The bell was then sounded again and objects began to move. In the first phase of each trial the objects moved together vertically upwards and behind the occluder, a transition which took approximately 1500ms. The objects then continued to move at the same rate following one leg of a Y-shaped trajectory behind the occluder and emerging on either the left or the right. The occlusion was accompanied by the sound of harp cascade. The objects were completely occluded for 3000ms. Finally the two objects moved to a resting position in the upper left or right when a final sound effect was played and the objects disappeared, either
3.3. EXPERIMENT 1B - S/D LEARNING IN 8-MONTH-OLDS

Figure 3.4: A schematic representation of the anticipatory looking experiment. The experiment was divided into three learning and one generalization block, each consisting of 16 trials. On each trial two items would appear together at the bottom of the screen, move up and behind an occluder before reappearing on either the left or right depending on an Same/Different rule. Shrinking away to a point for all objects on one side or shrinking and spinning for all objects on the other side. This took a further 1500ms. The eye-tracker marked a data file with timestamps for the start and end of each trial and the beginning and end of the occlusion period.

3.3.2. Results. Although 17 infants were tested, data from only 10 of these was included in the current analysis. In general the infants were engaged with the task and sat through a large number of learning trials and over half the generalization trials (for the infants retained for subsequent analysis, these averaged 43.2 learning trials and 5.2 generalization trials). There seemed to be no relation between the order in which the tasks were carried out and likelihood of baby not completing the experiment. Nevertheless, the analysis of the results was hampered by the lack of reliable data for all the infants. Of the original 17 infants, 2 were
excluded because they became extremely fussy early on, 3 were excluded because the eye-tracker recorded almost no data for them. (A setting on the Tobii turns off the 'Infant Add-on' which helps with tracking the infant head movements. This setting was occasionally reset when the test machine was restarted). Finally, 2 infants were excluded because of difficulties synchronizing the fixation data with the event data, making it impossible to accurately determine were occlusions began or ended. There were several reasons for this, poor calibration occasionally resulted in very sparse data or systematic biases in the data (for example, all points shifted downwards and rightwards). Additionally, even in the best cases there was fixation data for only about 60% of the time. Finally the later trials are likely to be less reliable as infants generally became less interested in the stimuli with repetition and not all infants contributed data to the generalization phase.

For this reason, it was decided to pool all data and not look at individual performance. (Although separate video recordings of the sessions were available, these are unsuitable for performing any analysis because of the sensitivity of the task to timing and the exact location in which the infants were looking.) Furthermore, to compensate for the variation in the amount and quality of data collected only relative measures of performance were considered. Two regions of interest (ROI) were defined in the top left and top right and the analysis compares fixations and saccades to the correct ROI relative to those to the incorrect ROI. (See Figure 3.5.)

There are three possible measures that could be investigated; whether the first fixation after occlusion was to the correct ROI, the proportion of fixation time during occlusion that was spent looking to the correct ROI or the latency of the first correct look after the object reappears. For both of the former measures a relative score can be used. It should be noted that the first correct look is in fact the first correct look for which there is data is not a completely deterministic measure (gaps in the data could reverse the effective result). However by accounting for looks to the incorrect side it is possible get the best proxy measure possible from the data available. For the first look correct fixation measure, each trial is scored as a success (1) if during occlusion there was a fixation in the correct ROI and there was no prior fixation in the incorrect region, conversely if the first recorded look was to the incorrect ROI that trial is scored as a failure (0) and otherwise the trial is marked
Figure 3.5: An example of all the fixation data collected from a single participant during the occluded phase of the experiment across all trials. The white boxes indicate the regions of interest for the analysis of results. Green points are fixations on trials where the stimuli will reappear on the right, blue points are fixations on trials where the stimuli will reappear on the left.

as (-) indicating insufficient data. For the relative proportion of time fixated for each trial the cumulative fixation times to the correct ROI were divided by the total fixation to both ROI’s during the occlusion phase. Although the infants’ latency to look to the correct location is perhaps the most interesting possible measure of anticipation and learning, it is unfortunately much more problematic to get a good measure of look latency with sparse or unreliable eye-tracking data. Large gaps in the data mean that creates substantial noise variation and inaccuracy in measures of latency and moreover there is no comparison measure for the correct and incorrect ROI’s by as in there was case of the first look measure. Therefore it is not possible to perform any meaningful analysis on latency.

The results organized by block are shown in Figure 3.6, with the corresponding significance levels shown in Table 3.1. The significance level were worked out as follows, within a given block, the average proportion score for each participant was
Figure 3.6: Summary of the proportion scores for all participants in Experiment 1b averaged by block. The blue line represents the proportion of looking to the correct region of interest (ROI) relative to the total looking to both correct and incorrect ROI’s during the occluded period. The red line represents the proportion of trials on which the first recorded fixation during occlusion was to correct ROI.

Table 3.1: Significance scores by participant by block for Experiment 1b. All values are for two-tailed, one-sample t-tested comparing mean proportion scores for each individual to the chance value of 0.5. Significant values are shown in bold.
calculated and then one sample t-tests of these values against the chance performance level of 0.5 were run. On all three learning blocks the performance was significantly better than chance. Infants learned from within the first block to associate the four training stimuli with the sides they would reappear. Likewise on the first two blocks the infants would on average look first to the correct side significantly more often than chance. However, on later trials, their first looks were not predictive, the distribution of first looks to correct and incorrect ROI not differing from chance. It is possible that this indicative of infants learning the time structure of the events. Objects always reappeared 3000ms after they were occluded so there is no need to look towards the ROI before that point. An analysis of saccades and latencies at the point of re-emergence might be able to confirm this hypothesis. Unfortunately, for reasons mentioned above this type of analysis has not been possible with this sparse data-set.

The sparseness of our data may also explain why there only appears to be a trend towards having generalized the S/D rule in the final block by this analysis. This result is based on only an average of 2.7 data points per infant for the generalization block, compared to 12.7, 10.9, 6.7 data points per infant for each of the learning blocks respectively. Thus, the analysis is not comparing like with like when running the grouped data t-test on the data in the generalization block. Even in the final learning block where there is much less data due to early terminations of the learning phase. Another way to analyse this data would be to treat each data-point as completely independent score. This is justifiable because each data point comes from a single trial from a single individual and independent measures are generally more conservative than repeated measures analyses. Therefore, the last learning block and the generalization block were reanalysis as group data. A two-tailed, one-sample t-tests was performed on the 67 relative proportion scores of the final learning block 3 to see if they were significantly above chance. This analysis confirmed that relative proportions were significantly above chance, t(66)=3.27, p<0.001 (two-tailed). Then the same analysis was performed on the 27 generalization trials, in both cases we test against the chance level of 0.5. Looked at in this way, the generalization data also showed a significant deviation from chance, t(26) = 1.91, p < 0.03 (two-tailed). This confirms that the infants were able to generalize
3.3. EXPERIMENT 1B - S/D LEARNING IN 8-MONTH-OLDS

Figure 3.7: Relative proportion of available fixation data where infants were looking to the correct ROI, averaged across all included participants and grouped by blocks of four trials. L1, ..., L12 are the learning trials and G1 and G2 are the generalization with novel shapes. Error bars show then 95% confidence intervals and note that the error bars are much larger in the later learning and generalization phase due to many fewer data points.

from the learning shapes (yellow circles and blue squares) to the novel shape and colour pairings (green triangles, etc.).

The fact that there is a significant learning effect present from the first block of 16 trials suggests that learning is quite rapid so a further set of analyses was performed grouping the data into smaller blocks of 4 trials. Figures 3.7 and 3.8 show the results organized in this way for the the relative proportion of time spent looking to the correct ROI and the proportion of first looks that were correct. In both cases, the 95% confidence intervals for each point are shown. For both proportion and first look data the learning data initially show a similar pattern; infants appear to rapidly learn to predict where the pairs will reappear and look first and longer to that location. Later on in learning, the infants continue to look longer to the anticipated location but no longer look there first, although the larger error bars in the later learning blocks and the generalization blocks are partly due to fewer data
One final type of analysis that is possible is to investigate if there are different response patterns for different trial types. During learning there were four possible trial types (two circles, two squares, a circle and a square, or a square and a circle). The first two representing the relation Same and the last two representing the relation Different. The question is whether infants respond differently to these two types of trials. To test this, a repeated measures 2x2 ANOVA was performed on the average proportion scores across the learning with Same/Different relation as one factor and identity of left hand object as the other factor. There was a highly significant main effect of S/D relation, $F(1,9) = 14.77$, $p<0.004$ but no other main effects of interactions. The results can be seen in Figure 3.9. The infants appear to learn to predict where the pairs of different items will reappear but are at chance when the items are the same as each other. The fact that performance is at chance on both the Same learning stimuli and at a comparable high level for both the
3.3. EXPERIMENT 1B - S/D LEARNING IN 8-MONTH-OLDS

Different learning stimuli suggests that the infants are using something about the Different property to make their discrimination. This is further supported by the fact that this pattern also holds in the generalization trials as well where the data were significantly above chance for the different pairs, $t=2.07$, d.f.=12, $p<0.03$ (one sample, two tailed t-test versus chance level of 0.5). See Figure 3.10.

3.3.3. Discussion. The results of this experiment suggest that 8 month old infants are able to use at least some aspect of Same/Different distinction on an instrumental learning task with simple geometric stimuli. It appears that the infants rapidly learn which side the different pairs will reappear and look longer to this location. The infants can generalise the rule they have learned, performing as well with 8 stimulus pairs with novel colours and shapes in the test phase. The infants’ first look after occlusion is initially predictive of where the objects will reappear but as they learn the timing of the reappearance they stop looking directly to the correct location. However, the fact that infants only appear to learn to predict the Different trials was unexpected and suggests that their performance may not be
directly attributable to ability to perceive the abstract S/D relation, but may be due to a more low level interpretation. Although the data were incomplete and difficulties with calibration means that there was reduced accuracy in the fixations, there were no systematic biases and by using relative measures and group data it was still possible to perform meaningful analyses.

One interesting aspect of the results is the very rapid learning. Infants were already above chance in the first block of 16 trials (and indeed by the second sub-block of 4). Learning appears faster than McMurray and Aslin (2004), Experiment 3 which used a very similar AEM design. In their learning phase infants received just 16 learning trials and averaged only 53% accuracy. Three factors might account for this improved learning, infants in our experiment were older (8 months vs 6 months), each trial only started when the infant was looking towards the lower middle of the screen ensuring that all infants saw (at least part of) all the training trials. Thirdly the stimuli moved more slowly across the screen giving infants more time per trial and included being fully occluded for 3000ms (vs 750ms) giving them more time to direct their attention to the correct location. Furthermore, this rapid learning can
be found in other studies, with an alternative anticipatory eye movement design. Sobel and Kirkham (2006, 2007) found that 8 and 5 month old infants were able to learn a spatial dependency after just 4 exposures to the critical training event.

In the later learning trials and in generalization infants’ first looks after occlusion were not predictive. This suggests that infants are learning the time course of the reappearance. The fact that each trial had the same sequence of sound effects could also provide scaffolding for the infants to anticipate the point at which the objects reappeared.

The fact that infants only learned the dependency for Different trials but were at chance with Same pairings indicates that infants are responding to something other than a truly abstract S/D concept. Like the pigeons in Blaisdell and Cook (2005), infants could be responding to lower level aspects of the stimuli, symmetry, contrast, etc. However, unlike Young and Wasserman (2002) the infants learned better with the more complex exemplars. One interesting follow-up might be to test infants using larger arrays of items to see if, as in Young and Wasserman (2002), they greater number of items increased their awareness of the redundancy. However, again that would moving away from a genuinely conceptual ability. One possible explanation might be that the different stimuli are more complex and they are more engaging. One might predict that if the rich photographic stimuli from Experiment 1a were used, infants would learn both S/D dependencies.

3.4. Experiment 2a - S/D Discrimination in 4-month-olds using Photographic Stimuli in a Habituation Task

The success of the infants in two very different tests of S/D learning suggests that by 8 months of age, infants possess the ability to perceive and utilize abstract relationships. Although, the partial success in only learning the Different dependency in Experiment 1b suggests this task was not utilizing some abstract relation simpler than S/D. Thus it is possible that these two experiments have found two different abilities. An obvious question is if these abilities are present in younger infants and if the same pattern of results is found. Caron and Caron (1981) found that infants as young as 12 weeks old dishabituated to Different items after being familiarized to items showing the Same relationship. (However by not running the converse
3.4. EXPERIMENT 2A - S/D DISCRIMINATION IN 4-MONTH-OLDS

Figure 3.11: Results from the Same/Different habituation experiment with 4 month old participants. In neither condition did the infants habituate to the pairs of items shown initially and did not show any subsequent increase in looking to the test items with novel pairwise relation.

3.4.1. Methods.

3.4.1.1. Participants. A total of 15 four month infants provided data (M = 124 days, range 117-129 days, 9 female, 6 male). A further 3 infants were excluded due to fussiness. Participants were selected from the CBCD database according to the same criteria as in Experiment 1a.

3.4.1.2. Stimuli and Procedure. The stimuli and procedures were identical to those in Experiment 1a.

3.4.2. Results. The results of Experiment 2a are summarized in Figure 3.11. The same analyses as for Experiment 1a were carried out. Initially a 3x2x2 mixed ANOVA was performed with trial type as the within subject variable and experimental condition (Same or Different items in habit set) and experiment order...
(Experiment 2a or 2b first) as between subject variables. There were no significant main effects or interactions. Another analysis with sex replacing experiment order as the final variable also showed no significant effects or interactions (all F’s < 1).

Since variations in age can have more substantial effect for younger infants, a final analysis was carried with trial type and experimental condition as within and between variables and age as a covariant. In this case, age was a significant covariant F(1,12) = 9.42, p < 0.01 but no other effects were significant. This was an overall positive correlation between age and looking time, older infants looking longer over all. This could be due to younger infants taking longer to get settled into the experiment or perhaps getting more restless towards the end. However, since no other age effects or interactions were significant it does not affect the experimental hypothesis.

3.4.3. **Discussion.** The 4 month old infants in Experiment 2a did not habituate and did not show the same discrimination as the 8 month old infants had done. They did not look longer at pairs of items exhibiting a novel Same/Different relationship to that which they had been familiarized with. One possibility is that the infants were potentially capable of making the discrimination but because they had not habituated and therefore were not liable to show a novelty preference at test. In practical terms it may be very difficult to test this hypothesis. The infants in this experiment had up to 19 habituation trials each of up to 20 seconds in length prior to the four test trials. The infants saw an average of 18 habit trials and did not habituate, either increasing the number of habituation trials or increasing the maximum length of individual trials might give the infants more time to process the stimuli but just as likely is that the experiment would then be too long and lose the participants completely.

3.5. **Experiment 2b - S/D Learning and Generalization in 4-month-olds using an Anticipatory Looking Paradigm**

Experiment 2b repeats the method of Experiment 1b with four month olds. Infants were tested with the same anticipatory looking paradigm using the same stimuli and experimental conditions as described in Experiment 1b. As before, infants took
part in both Experiment 2a and 2b in a single visit. The order in which they did the experiments being counterbalanced.

3.5.1. Methods.

3.5.1.1. Participants. A total of 9 four month infants provided data (M = 123 days, range 117-128 days, 6 female, 3 male). A further 8 infants were tested but were excluded due to fussiness (3), synchronization problems in the eye-tracking data (3) or very sparse data due to poor calibration (2). The same infants also took part in Experiment 2a.

3.5.1.2. Stimuli and Procedure. The stimuli and procedures were identical to those in Experiment 1b.

3.5.2. Results. The quantity and quality of the data were very similar to those in Experiment 1b. The 9 infants that contributed data saw an average of 37.7 learning trials and 4.9 test trials. The same analyses as for Experiment 1b were performed. Figure 3.12 summarizes the results. As before, one sample t-tests against chance were performed on the block by block data for both the relative looking and first look data. Only the relative looking for the first block was significantly above chance \( t = 1.86, \text{d.f.}=8, p<0.05 \) (two-tailed). As before there was much less data for the later blocks with 12.6, 10.6, 5.2 and 2.9 data points per participant for each of the three learning blocks and the test block respectively. Therefore, another group analysis was performed; the relative looking data from the final learning block and the generalization block were analysed treating each data point individually and performing one sample t-tests against chance level. In neither case was the average proportion score found to be different from the chance level of 0.5. This indicates that overall the infants have not learnt the rule nor been able to apply it to novel cases.

Although, it appears that, in this case, infants are not learning, they were significantly better than chance in the first block and showed a trend towards the correct side across the whole experiment. Therefore we once again perform an analysis with the data grouped in blocks of four trials to look at the possible learning in more detail. The results are plotted in Figure 3.13 but no clear pattern emerges.
3.5. EXPERIMENT 2B - S/D LEARNING IN 4-MONTH-OLDS

Figure 3.12: Summary of the proportion scores for all 4 month old participants in Experiment 2b averaged by block.

Figure 3.13: Relative proportion of available fixation data where 4 month old infants were looking to the correct ROI, averaged across all included participants and grouped by blocks of four trials. Error bars show 95% confidence intervals.
3.5. EXPERIMENT 2B - S/D LEARNING IN 4-MONTH-OLDS

Figure 3.14: Graph showing the different accuracy for during learning for the different pairs of stimuli for 4 month old infants in Experiment 2b. Infants respond at chance when the two items are the same as each other but are above chance at predicting the reappearance of the different pairs. Error bars show the 95% confidence intervals.

Finally, as before, we looked at the item by item learning for the different stimulus types presented. The relative looking to the correct ROI across all of the learning trials were grouped according to the objects in each pair. The results shown in Figure 3.14. The infants in this experiment show a similar pattern to that seen in Experiment 1b (Figure 3.9). The same repeated measures ANOVA (Pair relation x left-hand object identity) was performed no significant main effects or interactions were found although the main effect of S/D relation was trending in the same direction F(1,8) = 2.73, p<0.14. This indicates that effect was not as strong as with the 8-month-olds since when the effect was not significant at the individual level but the trend suggests that there is a continuity between the performance of the 8 and 4 month olds. Furthermore, when the data is analysed at the group level By grouping the trials into Same and Different trials, it can be seen that as with the older age group, the 4 month olds appear to be learning the different rule only. See Figure 3.10. Moreover, the 4-month-olds also appear to generalise with different pairings only. Running two-tailed, one-sample t-tests on their performance on the
different trials in the third block of learn trials showed a 67% accuracy rate which was significantly above chance performance level of 0.5, $t(26)=2.23$, $p<0.03$ whilst their generalization performance showed an 81% accuracy rate which was also significantly greater than chance, $t(13)=3.12$, $p<0.01$. Whereas infants were at chance in both cases for trials with Same pairs, (all $t<1$).

3.5.3. Discussion. The 4 month old infants in Experiment 2b did not show learning to the same extent as the 8 month old infants had done. But they do seem to show a qualitatively similar pattern of responding; They learn the contingency for different pairs but not for same pairs and they generalise this learning to novel shapes. The fact that such young infants respond in this fashion supports our finding with 8-month-olds. Taken together with the same infants failure on the S/D habituation task suggests success in this task may be mediated by something other than abstract same-different concept learning. We return to this point in the general discussion below.
3.6. General Discussion

The aim of this study was to investigate the emergence of abstract concept learning in young infants. The two item Same/Different (S/D) relation was chosen as an example of a very simple concept that had previously been extensively investigated in the animal cognition literature but to a much lesser extent with human infants. Two new experimental paradigms were developed that tested infants understanding and learning of the S/D concept in a passive habituation task and an active anticipatory eye movement (AEM) task with visual rich and simple stimuli respectively. Two groups of infants, aged 4 and 8 months, were tested to provide a developmental perspective. In the habituation task, it was found that 8-month-olds but not 4-month-olds were sensitive to the S/D relationship. In the AEM task, it was found that both 4- and 8-month-olds responded in a similar fashion; learning to anticipate the re-emergence of occluded Different paired shaped but performing at chance when the pair of shapes were the same as each other. Additionally, both age groups transferred their learning in the Different case to trials with novel coloured shapes. This pattern of responding is not consistent with a full awareness of the S/D relation but may be the result of some other feature of the stimuli.

3.6.1. Habituation with rich stimuli. In Experiments 1a & 2a infants were habituated to pairs of photographic images of either matching or non-matching inanimate objects drawn from a set of randomly chosen unrelated objects. They were then shown two further trials in which novel objects with a novel relation were presented. The results were very clear cut, with 8-month-olds succeeding and 4-month-olds failing at the task. In both conditions, 8-month-olds dishabituated to pictures showing the opposite relation to that which they had been familiarised with. There was no effect of scene complexity, meaning that infants did not look longer at displays with 2 dissimilar items relative to 2 identical items either during habituation or at test.

As previously discussed in Section 3.2.3, these results provide clear evidence that 8 month old infants are sensitive to the abstract relation of Same/Different. This gives stronger support for infants’ S/D discrimination ability than previous findings. It is more direct proof than the implicit conclusion concerning S/D awareness that may be drawn from 7.5 month olds success on artificial grammar tasks (Marcus
et al., 1999; Saffran et al., 2007). It is a more robust finding than those of Tyrrell et al. (1991) and Tyrrell et al. (1993) where only a small set of objects were used and paired preference paradigms meant that Same and Different pairs were simultaneously presented. In the present study infants saw a wide range of unrelated objects whose only common feature was the S/D relationship between them and they only saw a single pair of items at a time.

The design of this study also provides continuity with the findings of Wright and Katz (2006) concerning S/D learning in monkeys and pigeons. These results show infants are also able to make a S/D discrimination using photographic stimuli, even more rapidly and with smaller example sets. This continuity with the animal results and the strength of the effect in 8 month old infants suggests the mechanism is more general than previously believed. Note that the animal participants were rewarded for learning while infants discrimination is a looking time measure of spontaneous novelty preference. Moreover, the failure of 4 month old infants to do so suggests that this is unlikely that is a simple innate response. However, the 4-month-olds did not habituate and this may account for their lack of novelty preference. Wright and Katz (2006) found that increasing the number of exemplars improved performance of the pigeons and monkeys and a similar manipulation may lead to S/D discrimination in the younger infants. The current design is probably at the limit for of attention for the number of familiarisation trials for such a young age group. (Infants saw up to 21 familiarisation trials each of up to 20 seconds.) An alternative approach might be to have many more shorter trials and repetitions. Mareschal et al. (2005) found that a regime of rapid, repeated presentation (each trial 2 seconds long seen multiple times) led 4-month-olds to succeed in forming categories based on perceptual correlations that would otherwise not be found before 7 months old. A similar design may reveal a sensitivity of 4-month-olds to the S/D relation.

Seven month old infants can learn the AAB/ABA discrimination with synthesised syllables (Marcus et al., 1999) and with pictures of dogs and cats (Saffran et al., 2007) but do not learn with musical tones, timbres, animal sounds or colour shapes (Marcus et al., 2004). Saffran et al. (2007) suggest that stimuli must belong to a familiar category if infants are to notice the higher order relations between them. But, in this experiment, the stimuli did not form a single meaningful category
and many of the individual stimuli were very likely to be completely unfamiliar to the participants. Nevertheless, 8-month-olds were sensitive to the higher order S/D relation between the items. Therefore, an alternative explanation could be that infants succeed because the complexity of the speech and photographic stimuli make the identity relations more salient, since Same stimuli will match on a greater number of perceptual dimensions, emphasising their similarity.

Meanwhile, the relative salience of matching pairs of items may account for the surprising finding that there was no differences in overall looking time between the conditions at either age tested. It might be expected that the Different slides are intrinsically more interesting than the redundant Same slides and that, therefore, infants in the Different habituation condition would look longer overall, especially at the start of the experiment. This was not found to be the case. It may be that when confronted with two different items simultaneously infants have more individual features to look at but with two identical items side by side they are more inclined to make comparisons between the objects. An analysis of the patterns of looking in the two cases could possibly establish if there was such a difference. Such an analysis was not possible from the video recordings because of the difficulty of accurately determining exactly where the infants were looking. Similarly, although it was observed that four month old infants appeared to spend long fixated on one item at a time whilst the older were more inclined to switch back and forth comparing the objects, there was no way to accurately measure or quantify this. Therefore, it is not possible to investigate whether an such difference might account for the failure of 4-month-olds to notice the S/D relation.

A solution might be to rerun the study with an eyetracker recording fixations and saccades, so as to assess if there are differences in patterns of looking between the two conditions (pairs of matched or mismatched objects) and between the two ages. It might be expected that younger infants primarily fixate on a single object and make fewer comparisons between objects whereas older infants do compare the objects making more active saccades between the two objects. Moreover, it might be expected that the older infants show different patterns of saccades and fixations across the two conditions. In the Same condition, older infants might be making direct one-for-one comparisons of features to confirm / driven by fact that objects are identical, manifest in rapid switching between two objects. Whereas in Different
case, they may spend more time determining the nature of the individual objects dwelling first on one then the other.

Unfortunately, this type of analysis is not possible with limited video-recording data collected in Experiments 1a and 2a. This speculation on the limitations of interpretations available given the available data, highlight the more general problem that looking time is a relatively 'opaque' measure of performance. As Aslin (2007) concludes looking time measures ought to be complimented with converging measures of performance such as measures of heart rate (Richards and Casey, 1992) and pupil dilation (Sirois and Jackson, 2007) as a proxy for infant engagement or eyetracking techniques which can be more informative as to the 'microstructure' of infant looking.

3.6.2. Anticipatory eye movement with coloured geometric shapes.

In Experiments 1b & 2b an anticipatory eye movement paradigm was used to show that infants can rapidly learn to anticipate where a pair of geometrical shapes would reappear but only when the pair of objects were different from each other. They spent longer looking to the correct region of the screen while the objects were occluded. However, their first looks during occlusion were not predictive. The infants were also able to generalise their learning to novel shapes that also shared the different relation. But their performance was always at chance when the two objects were the same as each other. This was a surprising finding and one that does not have any clear explanation. Nevertheless, the effect appears to be real since a very similar pattern of responding was found in both 4 and 8 month old infants.

As previously discussed in Section 3.3.3, the results of Experiment 1b suggest that 8 month old infants are able to use at least some aspect of Same/Different distinction in this task. It appears that the infants rapidly learn which side the different pairs will reappear and look longer to this location. The infants can generalise the rule that they have learned to novel pairs of shapes. The infants’ first look after occlusion is initially predictive of where the objects will reappear but as they learn the timing of the reappearance they stop looking directly to the correct location. Each trial had an identical sequence of sound effects which could also provide scaffolding for the infants to anticipate the point at which the objects reappeared.
Eight month old infants learned rapidly and were already above chance in the first block of 16 trials. Learning was faster than McMurray and Aslin (2004), Experiment 3 which used a very similar AEM design. But the rate of learning was comparable with that found in other AEM experiments conducted by Sobel and Kirkham (2006; 2007), where 8 and 5 month old infants were able to learn a spatial dependency after just 4 exposures to the critical training event. Moreover, this experiment demonstrated methodological improvements over McMurray and Aslin’s (2004), Experiment 3. Firstly, the start of each trial was contingent on infant attending to the correct location on the screen ensuring that all infants saw (at least part of) all the training trials. Secondly, the stimuli moved more slowly across the screen and were fully occluded for 3000ms (vs 750ms) making it easier for the infants to track the objects and giving them more time to direct their attention to the correct location. Finally, the scaffolding sound effects and encouragement from the caregivers seemed to cause the infants to complete substantially more trials. All of which gave the infants greater opportunity for learning.

The fact that infants only appear to learn to predict the Different trials was unexpected and indicated that their performance is not directly attributable to an ability to perceive the abstract S/D relation. Instead, it may be due to the infants responding to lower level aspects of the stimuli such as symmetry, contrast, complexity, etc.. When Blaisdell and Cook (2005) demonstrated two item S/D learning with geometric shapes in pigeons they were careful to choose a set of colours and shapes for their stimuli that would be highly dissimilar so that it would not be possible to account for transfer in terms of the similarity between stimuli. They also point to converging evidence from other studies that supports the case for an abstract S/D in the pigeons by ruling out any other common perceptual property across experiments (e.g. Cook et al., 1997; Cook, 2002). However, in Experiments 1b and 2b an S/D interpretation is not appropriate since there is no learning of the Same cases and so alternative mechanisms must be considered.

One simple explanation that ought to be considered is the possibility that infants’ performance could be accounted for because they were exhibiting a systematic side bias, as was found in some infants in the experiments of McMurray and Aslin (2004). This can be discounted on two grounds. Firstly the side to which the Same and Different pairings would appear was counterbalanced across infants, for half the
infants. Different pairs would go to the left and Same pairs would go to the right, whilst for the remaining infants this dependency was reversed. Therefore, nothing in the geometry of the experimental set up (e.g. the layout of the testing room) could systematically influence infant attention to one side or another. Secondly, a side bias would lead to a below chance performance on trials where the unbiased side was incorrect (in this case the Same pairs trials). Such a bias was never seen in the data with performance on the Same trials being very close to the chance level of 50% in all cases. See Figures 3.10 and 3.14.

One manipulation that might increase infants awareness of the S/D properties might be to introduce the objects sequentially (as in Saffran et al., 2007). In the current experiment the two objects appear together, grow and shrink as a unit, move together and disappear simultaneously. This could lead the infants to treat them as a single gestalt whole and thus reduce their awareness of the fact that there are (from the experimenter’s perspective) two objects with some relationship between them. Although there is evidence that younger infants are less sensitive to some gestalt cues (Spelke et al., 1993) by four months common movement leads to a strong expectation of object unity in partially occluded objects (Johnson, 2001).

In Young and Wasserman (2002), pigeons learned S/D discriminations better with larger arrays of items that increased their awareness of the redundancy / dissimilarity contrast. Although those findings do not bear directly on the performance of the infants in this experiment, one related explanation might be that the Different stimuli are more complex with contrasting colours and they were more engaging leading the infants to attend and learn with these stimuli. Although this interpretation seems implausible from an adult perspective, only further experiment could rule out such an explanation of the infant behaviour. It is perhaps, even more surprising that neither group of infants learned the Same dependency when the learning phase used only two objects (the yellow circle and the blue square). With just four trial types in the learning phase, it might be expected that infants could learn a direct association between each pairing and its eventual reappearance. However, this did not happen with either age group and the fact that the identical pattern was found in the generalisation phase suggest that performance must be accounted for in terms of some kind of abstract feature of the stimuli. Future studies could systematically vary the common and distinct properties of the stimuli to investigate
this. For example, using arrays of multiple objects or a wider selection of objects in the learning trials or using photographic stimuli to investigate the issue of the infants’ engagement with the stimuli. Without further studies, it is not possible to give a clear explanation for the surprising findings of Experiments 1b and 2b. Nevertheless, the identical pattern of responding between ages and between learning and generalisation makes it likely that there is some abstract process underlying the infants behaviour in this task.

3.6.3. Infants’ understanding of Same/Different as an abstract concept? While the results of these experiments show that 8-month-olds (but not 4-month-olds) respond to the relation of Same and Different with photographic stimuli and that both 4- and 8-month-olds learn and generalise some property of Different geometric shapes, questions remains as to the extent to which these abilities can be said to be abstract and the extent to which they can be said to be conceptual. The two are not necessarily synonymous. These question are of interest to developmentalists and in the study of infancy they have been particularly brought into focus by the work of Jean Mandler. Her book *The Foundations of Mind*, subtitled *Origins of Conceptual Thought*, (Mandler, 2004), argues that there is a fundamental difference between the perceptual and the conceptual. The former referring to implicit perceptual sensorimotor knowledge while “concept refers to declarative knowledge about object kinds and events that is potentially accessible to conscious thought.” (p.4, Mandler, 2004). In particular she claims that concepts are not just complex associations of percepts but are a different mode of representation, contrary to the position of the ‘British empiricists’ Hume and Locke and more recently researchers such Quinn and Eimas (2000). Furthermore, Mandler would argue that although concepts are derived from percepts, an innate process of perceptual meaning analysis creates concepts based on perceptual input and a core set of foundational image-schemas. (See chapter 4, Mandler, 2004) A related perspective is found in Susan Gelman’s book *The Essential Child* (Gelman, 2003) which argues in support of the importance of Essentialism in young children’s conceptual understanding of the world. Essentialism is the notion that “categories have a underlying reality or true nature that one cannot observe directly but that gives an object its identity” (p.3, Gelman, 2003). Starting from Mandler’s perspective, then, there can be two interpretations of S/D discrimination abilities found in
these experiments. Either these abilities are due to the infants detecting perceptual regularities in the stimuli they encountered or could be a consequence of their awareness of a concept of Sameness (and also, by implication, Difference).

Since, in Experiment 1a, the stimuli are all unrelated, it is hard to point to the perceptual features that the stimuli have in common and which would contribute to a concept of sameness. Sameness appears to be a concept but not a category. Here, the commonality between instances is of a different kind to that found in object-based categories such as trees or dogs. Infants as young as 3 months can form such categories (Eimas and Quinn, 1994) as can pigeons (Herrnstein and Loveland, 1964). In such cases, although it may not be possible to determine exactly which features define the category, it is clear that they are perceptually grounded. Members of these categories share many similarities in terms of texture, colour, shape and spatial arrangement of features. Whilst all the Same pairs do share a horizontal translational symmetry, this is not a shared perceptual feature per se, but operates more like an abstract rule applicable to those cases. As such Mandler would perhaps identify it as an image-schema, in common with such notions as CONTAINMENT, SUPPORT or AGENCY. Nevertheless, Sameness or object-identity without being a concrete or abstract object-category is fundamentally a perceptual property and may contrary to Mandler’s view develop from perception alone. A parallel can be found in Quinn’s (2003) account of development of spatial relation categories in infancy. Quinn reviews evidence that infants under 1 year do have non-object based concepts of spatial relations such as above, below, between but that these develop in sophistication and abstraction over the first year consistent with a perceptually grounded account, where a process conceptual analysis is not necessary. The failure of 4-month-olds to respond to the S/D distinction with photographic stimuli indicates that this ability also develops. Although this by itself does not necessarily support either position.

The results of Experiments 1b and 2b are more in line with a purely perceptual interpretation. Infants are not applying an abstract Same/Different rule to the stimuli but are abstracting some perceptual feature to identify the Different stimuli and to learn to predict where these stimuli reappear and generalise this learning to new Different stimuli. This ability is largely unchanged across the two ages tested who show a very similar pattern of response. Infants of both ages seeming to
possess the same ability to discriminate and generalise only with Different stimuli. A surprising finding made more so when we compare this to the successes and failures of the same group of infants on the habituation tasks. A result which is problematic for Mandler’s dual process account; if eight month old infants possessed an abstract conceptual awareness of the S/D relation that allowed them to succeed on the habituation task, why is this concept not accessible in the AEM task? If the abilities are based on a perceptual analysis then this difference between the tasks can be explained in terms of the substantial perceptual differences between the stimuli used in the task. This interpretation leads to the testable predictions that using simpler stimuli in the habituation task might lead 4-month-olds to succeed on the S/D discrimination whilst using more complex stimuli in the AEM may induce the older infants to learn the dependencies for the Same stimuli.

Of course, there is a further problem with a conceptual interpretation to the habituation results with 8-month-olds in that monkeys and pigeons can also make S/D discriminations with photographs (Wright and Katz, 2006). Likewise, Blaisdell and Cook (2005) show that pigeons can discriminate S/D relations with coloured geometric shapes. In neither case would any researchers be likely to attribute these successes to the animals possessing human-like concepts but these results do prove problematic for Mandler’s account, where perceptual meaning analysis is required to form and utilise abstract concepts. Mandler (2004) is dismissive of Young and Wasserman’s (2002) results, claiming that these are purely abstract perceptual abilities but does not address the wider comparative literature where the accumulation of data across a range of paradigms (see Section 3.1.1) make clear that animals can go beyond the purely perceptual features of S/D stimuli. Those results parallel those found in these studies which suggests that there is a very clear similarity between the abilities of young infants and those found in other species that depend upon the experimental set-up and perceptual features available in a given task. The S/D concepts of both infants and animals appear to be firmly grounded in perception.

This chapter has looked in detail at a single example of abstract rule use in young infants while the previous chapter focused on what might have been a wholly associative learning process. The following two chapters will aim to compare the two types of processing more directly. In Chapter 4, two sources of category information
will be made available to the infant allowing stimuli to be grouped according to either a rule-based and a similarity-based classification. A set of test trials will put these sources of information into conflict to investigate the extent to which infants are aware of and utilise rules and similarity. In Chapter 5 rules and similarity will be studied by investigating how children and adults learn and sort categories that have either a rule-based or a similarity-based structure built into them.
CHAPTER 4

Unidimensional versus family resemblance category structure in infancy

4.1. Introduction

In this chapter we examine the relative importance of rules and similarity in category learning in infants. Recalling the discussion in 1.17.3, Smith and Kemler (1977) and Smith and Kemler Nelson (1984) showed that in spontaneous categorisation tasks with artificial stimuli adults typically categorised items according to a single stimulus dimension, whereas young children (4-8 yrs) were typically more holistic, grouping by overall similarity. This was termed the holistic-to-analytic-shift account (1989b; 1989a). However, an alternative differential-sensitivity account (Cook and Odom, 1992) suggests that young children do use single dimensions but do so inconsistently and with different relative saliences. It is not known how infants would categorize these types of stimuli. This is, in part, due to procedural limitations involved in infant testing. Using a habituation-dishabituation design we aim to discover if infants will spontaneously attend to the overall similarity or if in the presence of a single highly salient feature, they selectively habituate to this. We habituated infants with sets of items which share both a family resemblance and a uni-dimensional single common feature. We investigate what infants learn about categories when two sources of category information are available to them. By presenting a set of ‘switch’ items where the two sources of category information are put into conflict we can see which (if any) cause the infants to dishabituate. This allows us to infer which style of processing the infants favour. It is unknown how infants would spontaneously categorise this type of stimulus. It is unlikely that infants will respond analytically even though the youngest children tested appear to do so. Therefore, It might be expected that they would respond holistically. However, if one particular dimension was of a higher salience infants might fixate
on this and respond in a uni-dimensional fashion. We test 5 & 11-month-olds using rich digitally manipulated photographs and the artificial stimuli from Smith (1981). We hypothesise that in all cases infants will be holistic processors. But we first consider how other results from infant category learning could relate to this question.

4.1.1. Category Learning in Infancy. Various studies have shown that infants as young as 3 to 4 months can extract perceptual prototypes from a set of stimuli (Bomba and Siqueland, 1983; Quinn, 1987) but these studies only used relatively simple gestalt forms. Studies with more complex stimuli appear to show that infants are able to respond discriminate individual features and make use of overall feature similarity and that these abilities develop over the first year of life. Strauss (1979) found that 10 month old infants extracted a median prototype from a set of 14 schematic face stimuli that varied on 4 dimensions. Whereas the same age group in Sherman (1985) appeared to respond more to a modal test stimulus from 4 schematic faces varying on 3 dimensions. The former result appears to suggested that by infants can integrate multiple dimensions but the latter result suggests that they may just be frequency counting across separate dimensions. One possible interpretation could be that greater complexity of the Strauss (1979) experiment lead to a integrated similarity judgment whilst the simpler arrangement in Sherman (1985) may lead to features being encoded separately.

Studies by Barbara Younger and colleagues clarified the abilities of young infants to respond to multiple features simultaneously. In Younger (1985) 10-month-olds were tested with artificial animal stimuli which varied on 4 dimensions each at 5 levels, in one condition features varied independently and infants formed a single category whilst in second condition features covaried leading infants to form two distinct categories. Younger and Cohen (1986) found that 4-month-olds could discriminate sets of cartoon animals having novel features but that by 10 months old infants were also sensitive to novel combinations of previously encountered features. This suggests infants can respond both to multiple features and are sensitive to their correlations but does not predict how infants would treat a single highly diagnostic feature.
In a series of experiments Quinn and Eimas (1996) showed that with 3- and 4-month olds do sometimes attend to just one part of a stimulus. The infants selectively attended to features from the head and face region of various animal pictures to discriminate between dogs and cats. However, the relative complexity of the stimuli and high salience that faces have for infants (Morton and Johnson, 1991) mean that it would not be accurate to conclude that infants are attending to a single or small number of dimensions. French et al. (2004) showed that infants and auto-encoder neural networks can form a broad category of dogs that includes cats and a narrow category of cats that excludes dogs and that this can be explained by in terms of the containment relationships of the distributions of stimulus dimensions (e.g. ear separation, body length). Reversing the distribution relationships reversed the asymmetric exclusivity in the network and infants’ responses to pictures of dogs and cats. Nonetheless, although French et al. (2004) were able to show that 3.5 month old infants are sensitive to the distributions of features, their design varied many dimensions simultaneously and so does not determine if infants selectively attend to a single dimension or feature.

Thus it is an unanswered question how infants will respond to a single highly diagnostically feature in the presence of a broadly varying family resemblance category structure. By using the habituation-dishabituation paradigm this experiment aims to discover if infants will spontaneously attend to the overall similarity of a set of stimuli or if in the presence of a single highly salient feature, they selectively habituate to this. In part due to procedural limitations involved in infant testing only a few exemplars can be tested in any given experiment and the relationship between items and multiple feature dimensions is necessarily highly constrained. However, as we have seen child and adult experiments have also used small sample sizes and simple low dimensional stimuli (perhaps to their detriment) and so it ought to be possible to make a comparable infant study to the existing work on rules and similarity conducted with older participants. Using artificial stimuli allows the experimenter to control for prior knowledge and have precise control over all variations in the images presented to participants. However, these stimuli may be criticised on the same grounds, that they are not as rich as real world category exemplars and do not allow participants to capitalise on their existing category knowledge. For these reasons we will investigate both real and artificial stimuli.
Experiment 1 uses artificial stimuli based on those in Smith (1981). Whilst Experiment 2a & 2b use groups of real world photographs with highly salient artificial manipulations.

4.2. Experiment 1 - 11-month-olds with artificial stimuli from Smith (1981)

In this experiment we habituate 11 month old infants to a set of artificial stimuli, simple house-like compounds of plain geometric shapes varying in colour, size, shape and orientation, based on those in Smith (1981), Experiment 1, see Figure 4.1. During habituation three of the dimensions (house height, roof shape and house colour) vary randomly, their values sampled from a binomial distribution, while the final dimension, the orientation of internal arrow remains constant. At test, two types of test trial put the uni-dimensional and family resemblance characteristics into conflict, one item (novel family test) retains the same internal feature while taking on extreme values on all other dimensions, another (novel feature test) has novel arrow orientation but is average on all other dimensions. A final completely novel item tests for infants continued attention to the experiment.

4.2.1. Method.

4.2.1.1. Participants. This version of the experiment was run with a total of sixteen full-term infants (8 male, 8 female, mean age 343 days, range 314 – 376). A further four infants were excluded due to fussiness and two were lost due to equipment failure. Infants were recruited via the CBCD’s participant database. Parents were not paid for their infant’s participation but travel expenses were reimbursed.

4.2.1.2. Apparatus. A Macintosh G4 computer running OSX and MATLAB (version 7) was used to control the experiment. The experiment also made use of the Psychophysics Toolbox (version 3.0.8, Brainard, 1997; Pelli, 1997) The computer controlled the display of the experimental stimuli and calculated from experimenter key-presses when the habituation criterion had been reached. The computer was connected to 49cm colour monitor and a infra-red camera and screen titler. These were situated in separate sound-proofed testing booth.
Stimuli. The stimuli were adapted from Smith (1981). In the original experiment, figures were constructed out of coloured card and there were four independent dimensions that could take one of five values, previously rated as evenly spaced in psychological space by adult raters. Figure 4.2 shows the variation across all dimensions. For this experiment, the stimuli were presented on a computer screen but the same proportions were preserved and the screen colours were matched as closely as possible to the original Color-Aid hue codes (RVR, RV, VRV, V VBV) . The height, roof pitch and colour were randomly and independently varied across all five steps of each dimension with frequencies corresponding to a binomial distribution, (see Figure 4.2). The internal feature was the arrow which was 6.5 cm in length (3.7°). For this study the orientation of the arrow was kept constant as the internal single feature and was only presented at the two extreme orientations of 5° and 85° to the horizontal. Each image was between 15.3 cm and 24.0 cm in height on the screen (visual angle 9.1-14.6°) .

4.2.1.3. Procedure. Infants were seated on their care-giver’s lap, in a darkened room, at a distance of 95cm from the monitor screen. The care-giver was instructed not to interact with the infant and was unaware of the experimental hypothesis. A low-light video camera situated below the monitor filmed the infant’s reactions to the experiment.

An observer viewed the infant on a separate monitor in the main lab and started and finished each trial on the basis the infant’s attention by pressing a control key on the computer when the infant was attending to the stimuli. Each habituation trial was finished if the infant looked away for 0.5s, or had looked for a total of 15s.
Figure 4.2: Binomial frequency distribution for each variable dimension in Experiment 1a. Given dimensional values -2, -1, 0, 1 & 2, the features could take these [1, 4, 6, 4, 1] times respectively.

The computer timed each trial and calculated whether the habituation criterion had been met. The control software was written in MATLAB specifically for the current study by the present author. In order to ensure that habituation had occurred a strong habituation measure was chosen; when average looking time for last three trials was 50% of looking time averaged across the first three trials, with a minimum of 6 trials. If the criterion was not met the infant was shown a maximum of 12 habituation trials.

At test, the infant was presented with two test trials in a randomised counterbalanced order and there was a 30 second limit on how long the infant could look. In all cases the observer was unaware of which type of item the infant was observing. Between trials a set of random sound effects were played to keep infants’ attention on the screen. The test phase consisted of three test items. An item from the same category but with a contrasting internal feature (novel feature condition), an item from the other category that possessed the same internal feature as the habituation stimuli (novel family condition). The order of presentation of these two test items was pseudo-randomly counterbalanced between participants. Finally, a totally novel, highly interesting stimulus was presented to test for recovery of attention. Figure 4.4 illustrates the design used for the Experiment 1.

In the habituation phase infants were presented with a sequence of randomly selected items with the constant internal arrow orientation and all other features
4.2. EXPERIMENT 1 - 11-MONTH-OLDS WITH ARTIFICIAL STIMULI FROM SMITH (1982)

Figure 4.3: Illustration of how test stimuli are determined online for each participant. Each blue dot is dimensional representation of one habituation stimulus. The light blue circle represents the average test stimulus which is the mean of all the stimuli seen during habituation. The yellow circle represents the dimensions of the extreme test stimulus which is the most distant vertex of the rectangle that bounds all dimensions seen by infant. Participant has seen these extremal dimensions before but not in this combination. (The same mechanism works analogously in the three varying dimensions of the experiment.)

Figure 4.4: Design for Experiment 1a. Infants are habituated to items which varying binomially in height, colour and roof shape, all with a constant arrow orientation. At test they are presented with a feature conflict (average house new arrow orientation) and a category conflict (extreme house exemplar with old arrow orientation). Test trials were counter-balanced across infants and were followed by a totally novel picture to test for recovery of attention. See method sub-section on stimuli for explanation of the features and their manipulations.
4.2. EXPERIMENT 1 - 11-MONTH-OLDS WITH ARTIFICIAL STIMULI FROM SMITH (1981)

Varying independently according to their binomial distribution. At test, the infant would be presented with two test items in counter-balanced order and one totally novel highly salient final item to test for recovery of attention. The two critical test items were an average category/novel feature condition and an extremal category/original feature condition. Because the dimensions of the habituation stimuli were randomly binomially distributed and the total number of stimuli seen was determined by the habituation criterion, each infant saw a different set of stimuli. Therefore the appropriate test stimuli needed to be determined online during the experiment. Figure 4.3 illustrates this process. The average category/novel feature item was determined by combining a novel feature (arrow orientation) with the average of all the stimuli that infant had seen at habituation. To determine the extremal category/original feature item, the computer determined the highest and lowest values the infant had seen on each dimension. This gives the bounding rectangle the corners of which represent the combinations of all extreme individual dimensions for that infant. The algorithm picks the shape that is the combination most distant from previously seen exemplars. Superimposing the original arrow gives the second test condition. Order of these two test items was counterbalanced; half the infants saw an extremal category/original arrow test item first whilst the other half saw the average object/novel arrow orientation item first. All infants saw the rabbit as the final trial.

4.2.2. Results. Prior to full analysis, a second experimenter blind to the hypothesis double-coded a randomly selected 25% of the videos and her results gave a Pearson correlation of $r=0.975$, $p<0.001$ with the original coder. Eight infants were tested using a visual attention getter (a pulsing green circle appeared between trials). Eight were tested using an auditory attention-getter (a 'boing' noise marked the beginning of each trial.) A mixed 2x5 ANOVA (Attention-getter vs Trial type) showed no significant effect of attention-getter so data were collapsed across both cases for subsequent analysis. At test ten infants saw the average (arrow changed) stimuli first whilst six saw the extreme (arrow unchanged) stimuli first. Again, a mixed 2x5 ANOVA (Test order vs trial type) showed no significant effect of test trial ordering therefore, all data were pooled in subsequent analysis. Two separate ANOVA’s were conducted because of uneven cell sizes across the four attention getter by test order conditions.
Figure 4.5: Mean looking times in seconds for Experiment 1. There was a highly significant difference between average of the first three and last three habituation trials indicating that habituation had occurred. However, at test the infants only dishabituated to the novel test item.

Figure 4.5 shows mean looking times with confidence intervals for all 16 infants for the habituation and test trials. To test the experimental hypothesis that infants will habituate during familiarisation and only dishabituate to the novel family, same feature test trial and the completely novel trial, a series of one-tailed paired-samples t-tests were performed on the data. There was a highly significant difference $t(15)=5.70$, $p<0.001$ between the mean looking time across the first three habituation trials and across the last three habituation trials. This suggests that infants did habituate to the initial set of stimuli. There was no significant difference between the average of the last three trials and the average (arrow changed) trial condition or with the extreme (arrow same) condition suggesting that infants did not dishabituate to either stimulus. A comparison between last three habituation trials and the novel test stimulus (the bunny rabbit) was highly significant $t(15)=3.98$, $p<0.001$. The infants do dishabituate to the completely novel picture. Thus, the results appear to indicate that infants do not discriminate between either of the test stimuli and the habituation set. Because the test trials could last up to 30 seconds while the last habituation measure is actually an average of three 15 second trials it is not entirely appropriate to compare them directly (although it would lead to
4.2. EXPERIMENT 1 - 11-MONTH-OLDS WITH ARTIFICIAL STIMULI FROM SMITH (1981)

a more conservative estimate of any dishabituation effect than other measures). Therefore, we performed three other paired-sample 2-tailed t-tests comparing the three test trials with each other. This analysis shows the same pattern with no significant difference between the family resemblance and novel feature test trials t(15)<1, but that infants looked longer at the completely novel bunny compared to both the family resemblance t(15)=2.70, p<0.02 and the novel feature test items t(15)=3.57, p<0.003.

However, only 8 out 16 infants met the habituation criterion of a 50% drop in average looking time, the other 8 saw a full set of 12 habituation trials before the test phase. It is possible that those infants that did not meet the habituation criterion did not fully encode the initial stimulus. A further analysis was conducted on just the sub-group who habituated. The last three habit trials were significantly shorter than the first three (t=-3.8, d.f.=7, p<0.005) This was entailed by the habituation criterion. This group also dishabituated to the completely novel test stimulus (t=-3.04, d.f.=7, p<0.01) but they did not dishabituate to either the average or the extreme test stimulus.

4.2.3. Discussion. In Experiment 1a, the habituation stimuli were very simple geometric shapes that varied on three dimensions (height, roof shape and colour) and remained constant on a further internal feature, the orientation of an arrow. Thus the stimuli were relatively simple for the infants to encode and as a group they habituated to these stimuli. But they do not seem to dishabituate on the basis of either a single feature or extreme exemplars. Whilst readily reorienting when the completely novel bunny rabbit was presented. One reason might be that they are in fact grouping items by overall similarity and hence do not dishabituate to the change in the arrow orientation but also view the extremal example as not sufficiently different from the items seen at habituation. That is, they have formed a category that includes all the test items. Unfortunately, from our null finding we cannot make this conclusion but it is suggestive of possible follow up experiments.

This possibility is further increased by the fact that the infants only had a small number of examples from which to form any category. If in this case, the infants have formed a broad category that includes our extreme example, then the discrimination between items could be increased by increasing the contrast along
An additional complication is that the dimension themselves might not be equally salient to an infant. These stimuli were originally chosen because they had been used successfully with adults and young children (4+ yrs) by Smith (1981) With a control group of adults, Smith pre-tested discrimination along each dimension to ensure that each step was approximately equal in psychological space. There is little reason to suspect that infants would have the same reaction. In particular the colour dimension seems to have a narrow range of contrasts. Whilst overall height (for example) might have been very salient reducing the category to a set with only one main dimension. An improvement to the current study might present infants with a larger training set of more distinctive items with several equally salient but perceptually orthogonal dimensions. For example, avoiding the use of colour as a dimension or keeping the areas of each part constant whilst varying the shape. Pigeon researchers face similar problems when trying to design visual category example for the birds to learn. A collaboration with these researchers or use of their materials may prove beneficial.

An alternative approach would be to use far richer stimuli based on real world categories which have both a family resemblance and a highly salient uni-dimensional feature. This is the aim of Experiments 2A and 2B.

4.3. Experiment 2a - 11-month-olds with Photographic Stimuli

In this study, we aim to follow the design of Experiment 1 but using photographic stimuli. We chose pictures of cars and houses as two real world categories that would provide a range of possible stimuli suitable for the family resemblance comparison. They were isolated from their backgrounds to remove further variation and we chose to make digital manipulations to these items to give them all a single common feature which in each case would a highly distinctive surface texture. For the houses this was applied to the roofs and to the windscreen and side windows of the cars. Thus in each case the feature was defined by natural boundaries within the photograph and did not stand out but appeared to be an integral feature of the
4.3. EXPERIMENT 2A - 11-MONTH-OLDS WITH PHOTOGRAPHIC STIMULI

objects. Additionally, these areas were on the upper parts of the objects in approximately the same spatial locations within the objects and taking up approximately the same proportions of the total areas. We use a between subjects design where half the infants are habituated to cars and half to houses. At test we shall put the family resemblance and the single feature into conflict by showing infants one test item from the same family with a novel feature (i.e. a new pattern) and another test item from the contrasting family but with the same pattern as seen during habituation. We hypothesise that infants will dishabituate to the novel family test item.

4.3.1. Method.

4.3.1.1. Participants. A total of sixteen infants were tested (10 male, 6 female, mean age 330 days, range 316 – 344). A further 4 infants were excluded due to fussiness (2), experimenter error (1) or failure to attend to one of the test trials (1). Infants were recruited as before.

4.3.1.2. Apparatus. A Macintosh G4 computer running OSX and MATLAB (version 7) was used to control the experiment. The experiment also made use of the Psychophysics Toolbox (version 3.0.8) The computer controlled the display of the experimental stimuli and calculated from experimenter key-presses when the habituation criterion had been reached. The computer was connected to 49cm colour monitor and a infra-red camera and screen titler. These were situated in separate sound-proofed testing booth.

4.3.1.3. Stimuli. There were two sets of stimuli, photographs of cars and houses. There were a total of 26 individual original photographs (13 houses, 13 cars). In both cases the photographs were manipulated using the GNU Image Manipulation Program (GIMP version 2.2). Initially, each object was isolated from it’s original background and placed on plain mid-grey background. They were scaled to be approximately the same size, such that each image was around 20 cm in height on the screen (visual angle 11.9°). An additional photo of a bunny rabbit of a comparable size and on the same grey background was used as the completely novel test item. Further manipulations were performed to the car and house photos to produce the highly salient single feature that was comparable across the two categories. For
4.3. EXPERIMENT 2A - 11-MONTH-OLDS WITH PHOTOGRAPHIC STIMULI

Figure 4.6: Design for Experiment 2a. Infants are habituated to items from one category (e.g. houses) all with a single highly salient artificial feature (e.g. a leopard-print texture on the roof). At test they are presented with a feature conflict (novel feature) and a category conflict (novel item) trial counter-balanced across infants, followed by a totally novel picture to test for recovery of attention. See method sub-section on stimuli for explanation of the features and their manipulations.

Both sets of items, similar sized areas in the upper part of the object were selected and given a new artificial texture using tools in the GIMP package. Areas were chosen so that they were naturally defined by boundaries of the objects. Thus, for the houses, the roof of each one was given a new texture. Whilst for the cars, the main windscreen and side windows were re-textured. For each image, one new version was created using a leopard skin print pattern, another with a deep blue “swimming pool” effect.

Each participant was habituated and tested on either cars or houses using a randomly selected and ordered sub-sample of 12 items from that category, each with leopard print texture. The final item used in the test phase but with the bright blue texture. This will be referred to as the feature conflict item, because it comes from the same category but is different in the single common feature. A single item from the alternate category but also with the leopard skin print was used for the category conflict trial. Here the single feature is the same as at habituation.
4.3. EXPERIMENT 2A - 11-MONTH-OLDS WITH PHOTOGRAPHIC STIMULI

Figure 4.7: Mean looking times in seconds for Experiment 2a. There was no significant difference between average of the first three and last three habituation trials indicating that habituation did not take place. At test, infants looked longer at the out of category test item and the novel item.

but the overall category is different. The picture of the bunny rabbit was the final completely novel test item. Some examples are shown in Figure 4.6.

4.3.1.4. Procedure. The experimental procedures and habituation criteria were largely the same as in Experiment 1a. However, the selection and presentation of test stimuli was somewhat more different. As before the test phase consisted of three test items. The first two items were the conflict trials. One was an item from the same family but with a contrasting internal feature (novel feature condition), the other was an item from the contrasting family that possessed the same internal feature as the habituation stimuli (novel category condition). The order of presentation of these two test items was pseudo-randomly counterbalanced between participants. Finally, a totally novel, highly interesting stimulus was presented to test for recovery of attention. Figure 4.6 illustrates the design used for the real stimuli.

4.3.2. Results. The video recordings of the infants were used to code the length of looking time for each trial. Twenty-five percent of the videos were double coded by an second experimenter blind to the experimental hypothesis. These scores were highly correlated with the judgments of the original coder, $r=0.97$, $p<0.001$. 
At test all infants saw the novel feature test stimulus first followed by the novel item test, and finally the totally novel picture. But for half the infants the habituation set consisted of pictures of cars, the other infants were habituated to houses. Therefore, a 2x5 repeated measures ANOVA (habit set x test trial type) was conducted. The ANOVA’s showed a main effect of trial F(4,56)=2.676, p < 0.041 and a series of planned comparisons were performed. Figure 4.7 shows mean looking times and standard errors for all 16 infants for the habituation and test trials. A series of one-tailed paired-samples t-tests were performed on the data. There was no significant difference between the mean looking time across the first three habituation trials and across the last three habituation trials. Infants did not habituate to the initial set of stimuli. Indeed only 2 out of 16 infants met the habituation criterion of a 50% drop in looking time over the habituation phase. The other 14 infants saw the maximum 12 possible habit trials. Given that there was no habituation, it is somewhat problematic to compare the last habituation trials and the test trials. However, there was no significant difference between the average of the last three trials and the feature-change trials. A comparison between last three habituation trials and the category change (feature same) showed a significant one-tailed effect t(15)=2.46, p<0.014) The infants’ response to the completely novel stimulus (the bunny rabbit) was also significantly larger than the average of the last three trials t(15)=3.00, p<0.005.

One interpretation might therefore be that infants are looking longer at the items
which are not of the same overall category as they saw in the habituation phase. Thus if they had seen numerous cars, they would look longer at the novel house and the novel bunny rabbit. This would be consistent with what is know about infants abilities to form global categories by this age. However, the interpretation is further complicated by the fact that during the test phase, each trial could last up to 30 seconds. Therefore the comparison to the averaged final habit trials is comparing samples with different possible range. Since many infants were reaching the maximum on both habit and test trials this aspect of experimental design could potential confound any differences between habituation and test.

Planned comparisons looked at the separate performance of infants familiarised with cars or houses respectively, as shown in Figure 4.8. The infants who were familiarised with cars did habituate, t(7)=4.76, p<0.001 but at test only dishabituated to the totally novel stimulus t(7)=3.32, p <0.003. Whilst the infants who were familiarised with houses did not habituate during familiarisation t<1, but did look significantly longer at all three test conditions on test, t(7)=2.91, p<0.006 for novel feature test, t(7)=3.78, p<0.001 for novel family test and t(7)=2.28, p<0.019 for totally novel test respectively. (All t-tests one tailed.)

4.3.3. Discussion. In contrast to Experiment 1, here the stimuli were visually rich and so the infants may have been more actively engaged with the task and the stimuli. This is supported by the fact that taken as whole group infants did not habituated over familiarisation phase. Looking at the overall results from the test phase apparently suggests that infants were grouping the stimuli by family resemblance because infants look longer at the test trials featuring novel global categories, namely the novel family test and the completely novel bunny rabbit.

However, when the results were analysed, grouping the infants according to which image set they had been familiarised with a different picture emerges. (See Figure 4.8). Now it appears that infants initially familiarised with cars do habituate and do not dishabituate to either the (novel) car with the novel windscreen texture or to the novel house with the familiar texture. Whereas infants familiarised with houses did not habituate but do nevertheless look longer at both the (novel) house with the novel roof texture and to the novel car with familiar texture. No simple explanation for this pattern of results is immediately apparent. It could be that
cars are both a more familiar and more interesting category than houses for 11 month old infants. If this were the case it could account for the longer initial looking to the cars and the habituation in that condition, with the change in single feature test not noticed because infants treat objects in this familiar category in a holistic fashion. Whilst the short looking to the test house stimuli for is again explained by this being a less interesting stimuli in general. Similarly the fact that the houses were a less interesting but less familiar category could explain why they are initially less engaging for the infants but are less likely to be habituated to. Whereas the infants do notice the change in the single feature at test with the novel house because they treat this as a collection of features rather than an integrated category and can notice single feature changes (Younger and Cohen, 1986). This group may also look longer to the novel car and novel bunny stimuli because these are more intrinsically interesting for infants of this age or because they have more novel features.

An alternative explanation might be based on lower level features. For example, the cars have a wider range of colours and some consistent high contrast features (e.g. wheels and headlights) that may attract the infant attention, whereas the houses have a narrower range of colours and no set of obvious features that are perhaps not as consistent across exemplars (e.g. doors and windows are lower contrast and less visually consistent). This may lead to different patterns of responding on test. Needless to say, both of these explanations are very highly speculative and cannot seriously be offered without further investigation and controls.

4.4. Experiment 2b- 5-month-olds with Photographic Stimuli

By 11 months old, infants’ world knowledge would be such that they could be expected to possess global categories such as vehicles and buildings (Mandler, 2004). This means that they could appear to be holistic categorisers when in fact they are making use of their prior knowledge. In addition, the evidence from Younger and Cohen (1986) indicates that by this age infants are sensitive to violations of correlations between features. For example, those infants habituated to house may notice the co-occurrence of the distinctive roof and the presence of some other distinctive feature (e.g. angular shape) in the habituation set and dishabituate to the absence of either feature.
4.4. EXPERIMENT 2B- 5-MONTH-OLDS WITH PHOTOGRAPHIC STIMULI

Both these factors could influence the pattern of responding. Therefore, re-running the experiment with 5 month olds would provide a useful comparison group, at this age world knowledge is not yet entrenched and infants respond more selectively to individual features but cannot track feature correlations (Younger and Cohen, 1986). We keep all aspects of the experiment the same as Experiment 2a, so that a direct comparison may be made between the two age groups.

4.4.1. Method.

4.4.1.1. Participants. In this version of the experiment a total of sixteen infants were tested (7 male, 9 female, mean age 154 days, range 145 – 164). A further 7 infants were excluded due to fussiness (2) or failure to attend to one or more of the habituation and test trials (5). Infants were recruited as before.

4.4.1.2. Apparatus, Stimuli and Procedure. These were identical to Experiment 2a.

4.4.2. Results. An identical set of analyses were performed as in Experiment 2a. Figure 4.9 shows the group data for all 16 participants and indicates the significant one-tailed paired samples t-test comparison. There was a significant decrease in looking between the start and end of habituation t =2.68, d.f. = 15, p<0.008, indicating that habituation had occurred. Comparing the average of the last three habituation trials with each of the test conditions showed that 5 month old infants...
4.4. EXPERIMENT 2B- 5-MONTH-OLDS WITH PHOTOGRAPHIC STIMULI

Figure 4.10: Graph showing the combined results of Experiments 2A and 2B grouped by type of familiarisation trial. On the left, mean looking times for first and last three habituation trials. (Shown for comparison to test trials.) On the right, the interaction of habit set (cars vs houses) with test trail (Novel Feature, Novel Family, Totally Novel).

dishabituated to all three test conditions, t(15)=−2.01, p<0.031 for the category change, t(15)=−1.96, p<0.034 for the feature change and t(15)=−2.24, p<0.02 for the totally novel item.

An initial 2x5 repeated measures ANOVA (habit set vs trial type) showed a highly significant main effect of habituation set F(1,14) = 9.11, p<0.009 and a significant interaction with trial type F(4,56) =6.07, p<0.001. A series of planned comparisons confirmed that the pattern of results was similar to that found in Experiment 2a. The infants who were familiarised with to cars did habituate, t(7)=4.61, p<0.003 and did not look longer to any of the test cases all t’s<1.4. Whilst the infants who were familiarised with houses did not habituate during familiarisation t<1.1, but did look significantly longer at all three test conditions on test, t(7)=2.77, p<0.03 for novel feature test, t(7)=3.42, p<0.01 for novel family test and t(7)=2.77, p<0.03 for totally novel test respectively. (All t-tests one tailed. )

4.4.2.1. Comparing Experiments 2a and 2b. Since there appeared to be a similar pattern of results between the two age groups, a further analysis was performed pooling all the data. A (2x2x5) repeated-measures ANOVA was conducted with Age Group and Habituation set as the between subject measures and Trial Type as the repeated measure. Figure 4.10 summarises these pooled data. There was a highly significant main effect of Trial, F(4,112) =5.14, p <0.001 and of Habit Set F(1,28)=4.46, p<0.044. Additionally, there was a Trial x Habit Set interaction
4.4. EXPERIMENT 2B- 5-MONTH-OLDS WITH PHOTOGRAPHIC STIMULI

F(4,112)=5.27, p<0.001. However, there was no significant effects or interactions with the Age Group variable, suggesting the 5- and 11-month olds are responding in similar way.

4.4.3. Discussion. Experiment 2B showed a very similar pattern of results to Experiment 2A in a much younger age group. Five-month-olds who saw the cars at familiarisation tended to habituate but then did not look longer at either the test car with novel patterned windscreen (new feature), the novel house with familiar pattern (new family) or the totally novel bunny rabbit. The five-month-olds familiarised with houses did not habituate but did look longer at all three types of test trial.

Although an analysis comparing the two age groups found no effect of age, the consistent pattern of response is somewhat surprising. The infants in Experiment 2a were twice the age of infants in Experiment 2b. Previous studies have found considerable developmental change in infant categorisation abilities over the first year of life. Older infants are more sensitive to feature correlations (Younger, 1985; Younger and Cohen, 1986) and by eleven months old infants can be expected to form more differentiated categories (Younger and Fearing, 1999). Furthermore, there is an apparent shift from global to basic level of categorisation (Mandler and McDonough, 1993; Pauen, 2002b) and from more perceptually to more conceptually based categories (Quinn and Eimas, 1997; Mandler, 2004).

Likewise, the analysis combining the two age groups showed that the infants did not form any consistent categorical distinction between cars and houses in either experimental condition (familiarised with cars or with houses), even though evidence from other studies would lead to the expectation that both age groups are capable of making global categories distinctions (Behl-Chadha, 1996; Quinn and Johnson, 2000). Infants familiarised to cars did not dishabituate when they saw a house. Infants familiarised to houses did dishabituate to the car but also dishabituated to the house with the novel roof pattern. The continuity between ages and the lack of any consistent global categorisation suggest that the infants are primarily processing the stimuli in terms of their low level features. Thus, it might be the case that the artificial feature manipulation introduced in this experiment has disrupted the global, natural categories that these infants would be expected to form. Care was
taken to make the manipulations as ‘natural’ as possible. The leopard-print and the bright-blue effects were confined to naturally demarcated areas of the objects concerned (the windscreens and pitched-roofs) so as to make it look like they were a contiguous, integrated feature of the object. Whilst this interpretation is ‘obvious’ to adults, it cannot be assumed that infants would parse the images in the same way. To answer this question, an important control condition for future research would be to rerun the habituation task using the original, unmanipulated images to see if a global cars versus houses distinction is present.

Another useful control would be to discover if the infants have any initial, prior preferences for the car or house stimuli. As was discussed in Section 4.3.3, it is possible that infants’ real world knowledge, interest and familiarity with the categories of cars and houses could be having a differential effect on their performance. For example, it might be expected that the infants from London who participated in this study have considerable experience with cars and thus the pictures of cars to be a natural and interesting category for them. In contrast, the buildings were all American suburban houses that would be both unfamiliar and less liable to hold together as a coherent category. The contrast between cars and houses may also activate a difference in manipulability if the infant interpreted the cars as toys. It is known that graspability affects object processing in infancy (Kaufman et al., 2003) and any such difference in this case may lead to a difference in relative interest in the stimuli. Given these many unanswered questions, it is unclear how the results from these experiments should be accounted for.

4.5. General discussion

Two sets of experiments were conducted to investigate infants’ responses to sets of stimuli with both single dimension and family resemblance category structures. In Experiment 1, 11-month-olds were presented with artificial stimuli that contained a consistent single feature and a common overall family resemblance. The infants habituated over the initial trials. In the test phase, they did not dishabituate to either the family resemblance or single feature test item but did dishabituate to a completely novel bunny rabbit. This suggests that all habit and test items were treated as members of one category. The stimuli were all simple geometric compounds made of similar components. Therefore, it may be that the test manipulations
were too subtle for the infants or the stimuli too boring. We have no simple way of assessing the salience of the dimensions in this age group. Alternatively richer stimuli may engage the infants better. This was the aim of Experiment 2.

In Experiment 2A and 2B, 11- and 5-month-olds were presented with richer digitally-manipulated photographic stimuli, in which the test stimuli had either a novel family-resemblence group (e.g. cars instead of houses) or a novel single feature (e.g. a blue windscreen in place of a leopard-print windscreen). There was an interesting asymmetry, in that responses depended on the nature of the habituation set. Infants familiarised with houses do not habituate but do look longer to both family resemblance (novel category) and uni-dimensional (novel texture) test trials, whilst infants familiarised with cars do habituate and only looked longer at the completely novel test trial. This pattern was found consistently across both the age groups suggesting that this was a real effect caused by the difference between the familiarisation sets. Additionally, the results do not accord with any natural category structure even though both age groups would be expected to be able to make a global level discrimination between vehicles and buildings. This suggests that the digital manipulations did have an effect on the infants’ categorisation, albeit an effect that was manifested inconsistently between conditions and difficult to account for without further comparisons being run.

Categorisation is a flexible process even in infancy. Ribar et al. (2004) found that 6-, 10- and 13-month-olds could form narrow category of black and white animals or a broad category of land animals depending upon familiarisation, whilst Horst et al. (2005) found that function and appearance can alter whether infants process sets of items first by shared categorical features or initially as individual exemplars.

It may be that infants in the low complexity case of Experiment 1 are quickly able to form a categorical representation whilst infants in the richer case in Experiments 2a and 2b are responding to only individual exemplars and their features. Alternatively, French et al. (2004) showed, the distribution and relative salience of stimulus features can account for this asymmetrical category learning and feature sensitivity in infants. It is possible that the set of houses was more consistent and therefore easier to form into a usable category, allowing both family resemblance and single feature discrimination, whereas no clear category was formed for the cars and so no discrimination occurred. However, unlike French et al. (2004), which looked
at asymmetries between two very similar categories cats and dogs, it difficult to
determine exactly what set of features ought to be measured or compared when
contrasting cars and buildings. Finally, Quinn (2002) puts forward a theory that
familiarity influences infant category learning. If the stimuli are unfamiliar to the
infants will initially treat them as individual exemplars whereas with familiar mate-
rials they may form both a summary, categorical representation and learn individual
exemplars. This theory might predict that the familiarity of cars would allow those
infants who habituated to cars to from a categorical representation and notice the
single dimensional feature change. In fact, the exact opposite occurred; it appeared
that infants habituated to houses were the ones who formed a category (and hence
dishabituated to the novel car) and also noticed the single feature change (change
in the roof.) The answer to this question clearly interacts with prior knowledge
of the category in question and in this case can only speculate about the relative
familiarity of the materials. A related question from an adult perspective relates
to how one would expect a single consistent feature to be treated. Will a single
consistent feature across a group of stimuli be considered surprising because of its
consistency or will it be ignored because of its lack of variability?

Related to this is a question of whether the experimental set-up was appropriate
for our aim of drawing attention to the single salient dimension that we have ma-
ipulated. How are babies supposed to know that it is the consistencies rather than
the differences between stimuli that are important? This is an especially pressing
question in the case when each stimulus was rich in perceptual detail. Is it possi-
ble that increasing the salience of the single feature (e.g. make it brighter, larger,
flashing) would lead to a direct awareness of it. However, in the process, one could
lose the awareness of the object as a whole, with infants just aware of the bright
/ large / flashing aspect and treating everything else as background. In addition,
paired preference task might give a different result. Oakes and Ribar (2005) found
that simultaneous presentation of stimuli lead to different patterns of preference
as compared to a sequential presentation of stimuli. With each stimulus presented
individually the infants never directly see that the stimuli share a common feature.
It is possible that if each trial featured several items presented together then babies
attention is drawn to the matching patches of texture or the common direction of
the arrow.
Finally, these results do not resolve the debate between differing accounts of early rule-use in terms of a holistic-to-analytic shift (Smith, 1989b) or in terms of differential sensitivity to dimensions (Cook and Odom, 1992). The results from Experiment 1 showed that infants did appear to form a single holistic category for the simple geometric compounds presented but because only one unidimensional was tested, the possibility that infants would have been preferentially sensitive to one of the other dimensions cannot be ruled out. In any case, the intrinsically noisy nature of infant data would make it difficult to detect analytic responding. In contrast, the results of Experiments 2a and 2b, present the possibility that in some cases infants may track both family resemblance and single feature ‘rules’ while in other cases they do not notice either pattern.

The current chapter has investigated infants’ behaviour when two sources of category structure were placed in conflict with each other. It found that the different sets of stimuli lead to different types of categories being formed. This may have depended upon the distribution of features in the category sets or the relative interest and familiarity they held for the infants. In the next chapter, we ask similar question but this time with older post-linguistic children and adults, by comparing how children and adults learn and sort novel artificial stimuli which form families with either a rule- or a similarity-based category structure. The next chapter also presents a new method of investigation that addresses problems of ecological validity in the study of artificial category learning.
CHAPTER 5

Alien Rabble: A computer game architecture for assessing category learning.

5.1. Introduction

This chapter introduces a set of computer games created with the aim of taking the first steps towards two ambitious goals. Firstly, to bridge the gap between category learning in the laboratory and category learning in the real world. Secondly, to be able to test participants in a way that reveals the depth of their abilities and not the extent of their limitations, especially in the case when those participants are children. These two goals are related and complementary to each other. If participants are tested in a way that lets them perform in a laboratory setting that has some of the richness, complexity and engagement of the real world, then it is hoped that this will provide the best opportunity for them to demonstrate the rich and complex categorization skills they appear to demonstrate in their everyday lives. Conversely, if the aim is to test participants’ ability to learn or generate meaningful categories when confronted with highly multi-dimensional or naturalistic stimuli then it will be necessary to use large numbers of exemplars and a procedure that motivates the participants to remain engaged with the task. Furthermore, the depth of participants’ abilities will be best revealed by the ability to analyse their behaviour during learning in addition to the evaluation of their final classifications. It is hoped that "Alien Rabble", the computer game architecture presented here offers the means to achieve some of these ambitions.

The chapter begins by reviewing previous research which looks at differences in human category learning abilities, as they are found in the laboratory and the real world. This will show that some differences are the inevitable and even desirable consequences of constraints and controls that are imposed by experimental design, whilst others are at best unwanted side-effects of laboratory conditions and at worst genuine confounds of the abilities that are documented in more anthropological
settings. Additional emphasis is placed on the distinction as it relates to category learning in young children, showing that, in part, it is methodological limitations in the testing that may be responsible for the children’s seemingly limited abilities. The following section introduces the idea that suitably designed computer games might provide a means to address some of the short-comings of existing laboratory paradigms. It surveys some of the literature that makes the case that computer games provide a very rich and effective learning environment that might be particularly suited to engaging and assessing the highly computer literate children of a modern Western society. Since computer game design is not within the remit of a typical psychologist’s training and experience, Section 5.3 makes a brief digression into the technological details of the Alien Rabble system. This section is intended to only cover such details as are pertinent to the general aim of producing a framework for assessing category learning. It covers some of the principles of software design and development and outlines the capabilities and limitations of the chosen set of software tools and the current version of the game architecture. Next, two specific tasks are presented. The Alien Rabble system is used to conduct developmental studies of category learning, investigating the differences between children’s and adults’ approaches to rule and similarity-based classifications in supervised or unsupervised learning tasks. The theoretical contrast between rules and similarity is briefly reviewed and particular experimental procedures are described. The results of these studies are analysed and discussed before a final section surveys some future directions this work may take.

5.2. Ecological Validity in the study of Category Learning

There is a world of difference between the laboratory and the ‘real’ world, as scientists themselves are the first to appreciate. The problem of ecological validity concerns whether it is appropriate to extrapolate from an experimental result to a similar conclusion about the natural world. The closely related term external validity concerns whether an experimental finding is generalizable. The problem is of particular concern in the social and cognitive sciences where the measurement problem and the issue of demand characteristics are well known (Orne, 1962). Questions of ecological validity do not take away from the validity of laboratory findings themselves, as Mook (1983) famously articulated, external invalidity can
Table 5.1: Feature structure used in Medin et al. (1987) and numerous other category learning experiments. A family resemblance sort is possible, as shown, grouping the two prototypes A0 and B0 with the ‘one-aways’, items which are different from prototype in just one feature. Notice that uni-dimensional partitions on any one of the four dimensions are also possible.

be a good thing. Experiments are sometimes designed to explain the findings of other experiments and to show “what can happen” not necessarily “what does happen”. Likewise, Chow (1987) makes the point that ecological validity can rather be thought of as a feature of a theory rather than of an experiment. Whereupon the question becomes one about interpreting how a particular theory applies in the laboratory and the real world. Moreover, in this case, there is a strong meta-theoretical point that increasing the similarity of the laboratory setting to the real world will reduce the extent to which ecological validity is a concern in that instance.

5.2.1. Categorization inside and outside the laboratory. A recent edited volume produced in honour of Douglas Medin (Ahn et al., 2005), makes much of the fact that Medin has had two distinct strands to his career. Through careful experimentation with highly simplified materials and category structures, Medin investigated the exemplar versus prototype debate within category learning (Medin and Schwanenflugel, 1981; Medin and Smith, 1981; Medin et al., 1987) and also produced one of the first mathematical models of how categories are learned, the context model (Medin and Schaffer, 1978). Medin also took anthropological study of category expertise seriously. Together with colleagues, he investigated how different types of tree experts (biologists, landscapers and tree surgeons) formed different categorical structures and hierarchies (Medin et al., 1997; Lynch et al., 2000). In other studies they looked at the conceptual organisation of fish species by sports fishermen and Native Americans in Wisconsin (Medin et al., 2006). The particular conclusions of these studies are not so important here but the methods are worth contrasting.
The study by Medin et al. (1987) consisted of a series of seven experiments investigating category construction and family resemblance. The general procedure required participants to examine a set of drawings of stimuli or lists of features and divide them into two equal piles. In almost all cases the abstract structure of the stimulus sets was the same; each item varied on between four six binary dimensions and items were selected so that they could always be sorted according to a family resemblance structure or partitioned according to the value on any one dimension. The structure is shown in Table 5.1. For example, in Experiments 1 and 3 the stimuli were line drawings of hypothetical animals which varied in head shape (spiky or rounded), body pattern (spots or stripes), number of legs (4 or 8) and tail length (long or short). And thus the prototype A0 would be a spiky headed, spotted creature with 4 legs and a long tail. Each participant sorted just the 10 items, just once without any prior interaction with the materials or any other goal than to classify them. Sometimes participants received functional prompt, (e.g. to group creatures that might have good flying ability). The main finding was that family resemblance sorting was very hard to induce.

Medin et al. (2006) looked at the knowledge of expert fishermen from two groups, majority culture fishing club members and Menominee Native Americans. Each individual had over 20 years of fishing experience in the same area of Northern Wisconsin. Medin et al. sought to understand any differences between the groups in cultural terms, where “culture should be studied as a statistical distribution of mental representations and their public expressions among a given population in a given physical context.” (p. 245, ibid.) They produced a detailed background ethnographic sketch of their two groups and conducted two types of experiments. In Experiment 1 participants had to were given cards with the names of upto 44 native fish species and asked “put the fish that go together by nature into as many groups as you want.” (p.250, Medin et al., 2006) and were then asked to form larger super-clusters and smaller sub-groups, as they deemed appropriate, repeating the process as necessary to form a detailed hierarchy. Experiment 3 was essentially the same with different participants and the initial instruction to group by habitat. Experiments 2 and 4 presented all possible pairings of 21 species and probed the participants to list the interactions between A eats B, B eats spawn of A, etc. This took about 10 seconds per pair. Experiment 4 was the non-speeded version of the
same task with a smaller set of probe pairs. In all cases the experiments produced a
detailed set of knowledge per participant, including matrices of hierarchy relations
(Exp. 1 and 3) or interactions. Using a sophisticated analysis technique called
Cultural Consensus Model (Romney et al., 1986), Medin et al. looked for consensus
between the two cultural groups as well as within and between group differences.
The analysis was extensive but the general finding of the first two experiments
was that both groups shared a broad common pattern of classification somewhat
different from the strict biological taxonomy but that Menominee group generally
had more ecological and positive evaluations whilst sports-fishermen’s judgments
were influenced by the perceived prestige of the fish. In Experiments 3 and 4 there
were no group differences indicating that when given time or appropriately directed
focus both groups of experts did have access to world knowledge.

These two types of study could not be more different in their approach. The for-
mer testing spontaneous judgment of minimal importance to the participant with
very simple materials having predetermined highly arbitrary binary dimensional
structure. The latter performs very complex mathematical analysis on the detailed
and open-ended probing of the complex real-world knowledge of highly experienced
experts. Almost the only thing they have in common is that they are both are stud-
ies of concept learning (and were both conducted by Doug Medin). Reassuringly
both do find a consistent pattern of responding across participants. So even if they
are not testing the same thing they are each to themselves internally valid. Greg
Murphy (2005), in his chapter in the volume already mentioned (Ahn et al., 2005),
also examines this contrast. He reiterates that the ultimate goal is understand real
concepts and how they are really learned. In the light of the later anthropologi-
cally inspired work of Medin and colleagues and the earlier foundational work of
Eleanor Rosch (Rosch, 1975; Rosch et al., 1976), the artificial materials and un-
natural category structures of earlier work of Medin and other laboratory studies
seem more distant from the original goal. In an earlier book chapter, Murphy
(2003) also looked at the question of ecological validity in category learning. He
reiterates the point of Chow (1987) that it is ultimately the theories that must
answer questions about ecological validity rather than the data. Nevertheless, he
considers that one of primary reasons for the continuing disputes concerning the
exemplar versus prototype debate, the role of prior knowledge and the contrast
between 'fast mapping' children (Carey, 1978) and slow learning adults (e.g., Medin and Schwanenflugel, 1981) is due to the focus on potentially misleading laboratory results. Strangely, Murphy (2003) does not call for greater ecological validity in category learning experiments per se. Rather, he suggests that laboratory work can remain complementary to other more naturalistic means of enquiry such as the psychoanthropological work of Medin and colleagues or studies of everyday word usage. Additionally, a process of ecological analysis to be applied to change the emphasis from a question of “which theory is correct to when each theory is correct” (p. 37, Murphy, 2003, emphasis in original). This may re-frame the debate and change the emphasis from that focused on in current laboratory learning paradigms.

5.2.2. Category Learning and Category Use. Markman and Ross (2003) are more forthright in suggesting that “the richness of use of natural categories requires an extension of laboratory techniques for studying category acquisition” (p. 592, Markman and Ross, 2003). They contend that it is insufficient to just look at category learning if real world category knowledge is to be understood. One of their key points is that real categories are used in a variety of ways besides classification. For example, for making inductions, to communicate efficiently and as means of producing novelty. Moreover, Markman and Ross feel it is likely that categories are also learnt in a variety of ways that reflects the context in which they will be used. As such they feel one of the biggest problems with laboratory research in category learning has been the almost exclusive focus on classification. They provide a detailed analysis of difference between two widely used methods, contrasting category learning with property inference. They then highlight the potential of studies that combine classification and inference and studies where problem solving influences what is learned about categories.

In the section of their review comparing category learning with property inference, Markman and Ross (2003) emphasise that formally these two methods are completely equivalent. In the inference case a missing feature must be deduced from the category label and other information {Category label, feature 1, feature 2, feature 3, ?} whilst in the category learning situation the category label is to be deduced from a set of features {?}, feature 1, feature 2, feature 3, feature 4}. There is no logical reason why the category label cannot just be considered to be another
feature and yet empirical evidence suggests these two situations are psychologically very different. It appears that in classification tasks participants tend to focus on a small subset of diagnostic features that can distinguish between categories whilst in inference tasks they appear to focus on features that typical within a single category (Nosofsky et al., 1994; Yamauchi and Markman, 2000). Additional evidence comes from Susan Gelman and Ellen Markman’s work showing that children make special use of category labels (Gelman and Markman, 1986 and see below).

Markman and Ross feel that learning category structure is rarely a goal in itself and are often incidental to solving other problems and thus experimental approaches should reflect how categories are actually used. They describe Experiment 2 in a study by Yamauchi and Markman (1998) where the order of an inference and a classification task was manipulated. Participants who performed the inference first then the classification learned faster than participants presented with the tasks in the opposite order. This was interpreted as evidence that the first group of participants were acquiring a prototypical category during the inference task which facilitated their subsequent category learning, whilst participants performing category learning first acquired the diagnostic features of the category which do not facilitate their subsequent inference. In a somewhat involved problem solving task, Ross (1997) found that non-diagnostic categorical features encountered in course of solving an arithmetic problem were used in later classification. Participants had to ‘decode’ a numerical message from spy A or spy B according to two distinct arithmetic procedures, the first procedure drew attention to feature a (digit 2 multiplied by digit 5 always equalled 12) whilst the second drew attention to a different feature b. All the training messages contained both features a and b so these feature by themselves were not initially predictive but because they were operationalised as features associated with class A or B, they were used that way in subsequent classification task by a significant proportion of the participants.

Markman and Ross conclude their survey with three suggestions and three warnings for researchers. They suggest (1) that laboratory studies of category learning should widen the range of tasks that are used with a greater focus on category use, (2) that greater attention should be paid to the information processing demands of the particular task being considered and (3) that greater consideration should be given to a wider range of aspects of natural categories and the tasks they are used for.
Table 5.2: Comparison of the typical features of category learning as encountered in the laboratory and in the real world.

They caution that this is a very large problem, that inference is liable to be easier to investigate than other aspects of category use and finally that this diversity must be approached conservatively and methodically so that the variation can be mapped from task to task and the field is not faced with a plethora of models each applicable only to the results of a single task. These issues are revisited below. This section concludes by presenting a table which summarises some of the key difference between category learning in the laboratory and the real world (Table 5.2).

5.2.3. Assessing category learning in children. If there is an acknowledged problem with adult laboratory tasks not assessing the full range of adults’ category learning and knowledge, the unacknowledged problem with the restricted methods available to developmentalists is undoubtedly worse. Children’s lesser cognitive capacities, shorter attention spans, limited experience with the world and their restricted meta-cognitive abilities means that experimental designs must necessarily be greatly simplified. No researcher would be foolish enough to attempt to get a young child to sit through 40 blocks of learning trials and even if they succeeded, the chances of the results revealing anything other than boredom are remote. This section very briefly describes two typical category learning tasks used with children, triadic classification and word learning studies and examines the limitations inherent in the data that they provide and the conclusions that can be
5.2. ECOLOGICAL VALIDITY IN THE STUDY OF CATEGORY LEARNING

drawn. In particular, it is emphasised that neither of these types of tasks is directly
investing the learning of new categories.

For over 25 years, the triadic classification task has been the dominant paradigm in
the investigation of children’s categorization abilities at both the perceptual (Smith
and Kemler, 1977; Thompson, 1994) and conceptual level (Gelman and Markman,
1986; Imai et al., 1994). In all cases the basic procedure is the same, children are
shown a target item A and two possible comparison items B and C. Item B will
share some relation with A while C has a contrasting relationship. For example,
in Thompson (1994) the stimuli were coloured discs chosen so that that A and B
matched exactly on a single dimension of size or brightness while A and C were
similar on both colour and brightness. Whereas, in Gelman and Markman (1986)
the stimuli were line drawings of objects that either had a close perceptual match
or shared a common label. For example, A was an gray shark, B was a brightly
coloured tropical fish and C was a gray dolphin. When children were provided with
information about the fish (breathes underwater) and the dolphin (breathes at the
surface) and were told the shark was a fish and asked how it would breathe.

As these two examples illustrate the triadic classification task is used across a very
wide range on contexts but in all cases there are limitations. When assessing uni-
dimensional versus multidimensional grouping strategies as in the experiments of
Smith and Kemler, 1977 and (Thompson, 1994), the stimuli are very simple of
low dimensional complexity and only ever varying in two dimensions at a time.
The design nevertheless requires multiple comparisons to be made across many
conditions. In Thompson (1994) there were 6 potential comparison types (see
Figure 5.1 in Chapter 5.) and 10 questions per comparison per type. These were
analysed by counting the instances where participants appeared to be engaged
in certain types of rule-like responding. Raijmakers et al. (2004) criticise the rule
analysis of Thompson (1994) on the grounds that the strategies searched for and the
80% accuracy cut-off were arbitrarily chosen. Raijmakers et al. (2004) favour Latent
Class Analysis as a more sophisticated and statistical well-principled analysis. The
downside of this approach is that of even larger samples are required. For example,
in Experiment 1 of Raijmakers et al. (2004) 226 children were tested whilst in their
second experiment required a total of 357 participants. Perhaps more critical is
the fact that in none of these experiments are children creating, learning or even
using categories. What is being assessed is their spontaneous preference for unidimensional or family-ressemblance sorting.

Likewise children do not learn any new categories in the more conceptual studies like those of Gelman and Markman (1986). Here they are being tested on their accumulated world knowledge of conceptual hierarchies and evidence is being sought that children are Essentialists (Gelman, 2003), that they attribute a hidden essence (e.g. fishiness) to items that belong to the same family, even in the face of perceptual differences and in the presence of perceptually similar foils and can use this knowledge to infer non-obvious common properties of members of the same class (e.g. fish breathe under water). Gelman and Markman conducted lots of controls and tested children with living and non-living natural kinds (e.g. a pile of brown sugar compared with a pile of sand and some white sugar cubes) and their findings have been replicated but whilst this explains something about the structure of existing category knowledge it does not examine how that knowledge is acquired and results can sometimes show the opposite pattern. For example, in Imai et al. (1994) children favoured extending a new label on the basis of perceptual similarity (e.g. an apple and a red balloon) rather than taxonomic similarity (e.g. an apple and a banana). For this reason, Namy and Gentner (2002) criticise the triadic test on the grounds that only a single exemplar is presented for comparison on each trial. They modified the task so that two items (e.g. an apple and a pear) were initially presented and this increased the taxonomic responding, which they argued was due to the increased ability to compare exemplars. However, even if children favour one classification or another, this is not necessarily evidence that they cannot make or do not understand the alternative classifications, but rather a shortcoming of the task itself. For example, experiments conducted by Murphy (2001) showed that adults favoured a thematic classification over a taxonomic one when there was a strong thematic link between the probe and the thematic choice (e.g. bee, fly, honey). Yet clearly the adults understood both links between the test items.

Much of the evidence about children’s category learning actually comes from research into language learning and concerns the special status of labels and how they are acquired. Some research looks at children’s vocabulary and word usage as evidence for their understanding of category structure, particularly that of Sandra Waxman and colleagues (e.g. Waxman and Gelman, 1986; Waxman and Kosowski,
In his book *What video games have to teach us about learning and literacy*, James Paul Gee (2003), a professor of education, examines how in order to be successful good video games must challenge their players with tasks that are always on the edge of their capabilities and teach their players the new skills they will need to progress to the next level. He considers that good games are designed to exploit and encourage the skills and aptitudes that are indicative of good learning and they do so without ever being boring. His interest in the topic arose out of his experience of attempting to teach his own four year old son how to play his educational games, only to fail and effectively find his son teaching him. But no-one had taught his son, Gee realised that his son’s computer games were designed in such a way as to facilitate learning and discovery though exploration. They lead players gradually into the logic of the game and provide players with the opportunity and motivation to hone and practice a continually advancing set of skills. As an educationalist he realised that good game designers solved the problem of getting people to learn hard and challenging things. In fact, Gee (2003) manages to find 36 learning principles which are built into good games. Gee was primarily concerned with long, involved action adventure games.
that had a story such as *Tomb Raider, Half Life* or even the game that sparked his own interest, *Pyjama Sam in No Need to Hide When It’s Dark Outside*. As such many of his principles address the extended story-telling and discourse-building aspects of these games and so not all of his principles are relevant to a simple set of categorisation tasks disguised as a computer game. Nevertheless, a number of them are directly relevant and so they are quoted at length here (Appendix, Gee, 2003, pp 207-212):

1. **Active, Critical Learning Principle** - All aspects of the learning environment ... are set up to encourage active and critical, not passive, learning.

3. **Semiotic Principle** - Learning about and coming to appreciate interrelations within and across multiple sign systems ...

7. **Committed Learning Principle** - Learners participate in an extended engagement (lots of effort and practice) as extensions of their real-world identities in relation to a virtual identity to which they feel some commitment and a virtual world that they find compelling.

12. **Practice Principle** - Learners get lots and lots of practice in a context where the practice is not boring.... They spend lots of time on task.

17. **Situated Meaning Principle** - The meanings of signs (words, actions, objects, artifacts, symbols, texts, etc.) are situated in embodied experience. Meanings are not general or decontextualized. ...

22. **Intuitive Knowledge Principle** - Intuitive or tacit knowledge built up in repeated practice and experience... counts a great deal ...

28. **Discovery Principle** - Overt telling is kept to a well-thought out minimum, all owing ample opportunity for the learner to experiment and make discoveries.

All seven of these principles go towards the goal of creating a more ecologically valid approach to testing category learning in children. There are many other good reasons for choosing this approach. Using a modern video game paradigm one can create a visually rich environment and if necessary creating more natural life-like stimuli and one can test larger sample space with more exemplars. Artificial categories usually have very low dimensionality with heavily overlapping feature sets whereas our experience of real world categories is of widely separated groups in
a highly multidimensional conceptual space. Our experience of real world categories is the accumulated result of many, many encounters with countless exemplars. (Just think how many cars a typical urban child will have seen by his or her first birthday.) It is a problem of existing approaches to category learning with artificial stimuli that they use very small sets (around 10) items each of which varies in a very constrained way in very limited number of dimensions (often binary changes with just four or five features in the object.) Moreover, all participants are tested on the same materials. Partly this is done to control the task but it is also a consequence of the impracticality of creating much larger range of samples and producing the appropriate set of test cards for each case. Within a computer paradigm the stimuli can be constructed algorithmically allowing the game to automatically construct a set of exemplars to meet any of a number of conditional requirements. The control over stimulus creation also allows for genuinely scalar variation on a large number of dimensions, random noise introduced into training sets and participants may tested repeatedly with different materials each time. Furthermore, given the dynamic nature of a video game setting, wider range of properties can be included (e.g. auditory features, characteristic types of movement, causal properties, possibly even intentional states). Since the objects are presented in a situation the role of context can be investigated.

Games also provide a mentally engaging task, one that does not have to be primarily about classification. Hence they can provide a means of meeting Markman and Ross’s (2003) criteria that category learning should be investigated in situations where categories are used for something besides classification. Since the game setting can disguise the nature of the investigation it may also remove some of the demand characteristics that are found in any traditional laboratory study. Framing the test as a computer game provides a natural explanation for the task that is less artificial, and where attention can be drawn away from experimenter’s objective so that category learning may be incidental. Furthermore, a stimulating game with genuine problem-solving elements will have the participant cognitively engaged and will hopefully keep them engaged over a longer time span. This is particularly of concern when testing younger children. The computer interface also allows for making the experiment into a largely non-verbal task controlling for the linguistic sophistication of participants. Finally, the computer based nature of the task makes
it easy to administer and to control whilst at the same time making it possible to
collect a much richer data set than traditional tasks. As well as classification
judgments and the sorting of objects, the computer can record reaction times for
each of these judgments, track the order that objects were sorted in (are more
typical exemplars chosen first?) and even lower level information (the exact amount
of time that the participant could see any given object). All these reasons make a
strong case for investigating the plausibility of building a computer-based category
learning task.

5.3. Technical Background

In contrast to the previous section, which examined the benefits of game-based
approach to the study of concept learning, this section is more general. First, it
elaborates a few overarching principles that apply to all software design but that are
particularly relevant in this context. Then, in order to elaborate the future potential
of the system, it goes into some of the technical details of actual implementation
of the Alien Rabble system. This whole section can comfortably be skipped by
the reader interested primarily in the experimental content of the chapter which
resumes in Section 5.4.

5.3.1. Software design principles.

Iteration. Version 1.0 of any piece of software is never the first version of that
software. It more usually marks the end of the beginning. The first version the
developers are willing to release into the wild. The milestone version in a long
sequence of refinements that have lead up to that point. But equally it is never the
definitive version either. Software design is a highly iterative process. The develop-
ment of any new procedure, programme or system goes through many phases of
prototyping, testing and refinement. This is a consequence of the complexity of the
problem; the requirement to precisely specify each logical step of the process, the
complexities of multiple interacting subroutines and the many and various restric-
tions that come from operating on particular systems, with particular programming
languages and whilst using pre-existing tools, frameworks and libraries of functions.
Specifications can be given at the outset but software is a highly constrained prob-
lem space. Certain solutions will not be possible but this does not become apparent
until they are attempted. Whereupon a work-around, an alternative solution or a totally different approach is needed. Iteration is the key to progress. This was a key principle that the present author found in his 8 years as a professional developer: It is only ever many small steps that add up to giant leaps.

Modularity. The second highly relevant principle of modern software design is that of modularity, which is closely associated with the principles of code simplicity and code reuse. The idea of modularity is, of course, highly familiar to cognitive scientists and, in the realm of computer science at least, Fodor’s (1983) definition of modules do serve as a useful guideline. Efficiency, encapsulation and functional specialism are all desirable properties of software modules and the principle of modularity itself is a desirable goal when writing software. By decomposing the system into components, one simplifies the problems that must be solved and makes it possible to tackle each one in isolation from the rest. The modules that achieve this in modern ‘object-oriented’ programming languages are called classes. Unfortunately for Fodor, modern ‘object-oriented’ languages also make use of hierarchical and polymorphic structures where classes inherit their basic functioning from ever more abstract parents and adapt it their own needs. For example, in a computer racing game the classes for Ford, Ferrari and Fiat will all inherit and adapt a set of common logic from a more general car class and this in turn shares some properties with ambulance, fire-truck that were all inherited from vehicle and this in turn from the completely abstract class of entity. Again the advantage is simplicity, efficiency and code reuse. A programmer need only write one collision detection routine at the level of the entity class that can be utilised by all its children alike, Ford, Ferrari and even pedestrian.

Cognitive scientists will also recognise modularity in a second sense when they consider principles of good experimental design. Conditions are controlled by varying one aspect at a time. The inherent modularity of computer game software coincides well with the demands of an experimental psychologist. For example, if it is desired that a set of distractor items is added, this can be achieved without changing the operation of any other aspects of the game. Likewise, a countdown timer could be added in order to put the participant under pressure. By following the principle of modularity, a software designer can greatly simplify the problems to be tackled and divide their solution into manageable chunks. The reuse of solutions and benefits
of hierarchical structure mean that problems can be solved with efficiency and the imposed modular structure lends itself well to innovation and meet a scientists need for multiple, independent varying conditions.

**Reliability.** Once written and compiled, computer code is fixed and will execute in exactly the same fashion every single time it is called. Each object will look the same, each routine will follow the same steps, no protocol will be forgotten and all the data will always be recorded automatically. Even randomness can be controlled, the pseudorandom number generators can be started from exactly the 'seed' value each time so that every raindrop falls exactly the same way each time you simulate a thunderstorm. Nevertheless, although computers are completely logical it is remarkable how often they do get things wrong. In fact, it is their logical precision that causes their brittleness and propensity to break down. The fact they are programmed by humans also contributes to their downfall. Therefore, knowing oneself, one works with the expectation that one will make mistakes and that other unanticipated things will go wrong.

In part, this is done by building in a system of error catching and handling routines that ensure the system copes gracefully with unexpected events. Continuing on to the next instruction, in the case of minor glitches and halting with a clear and detailed log of the problem if there really is a problem. Substantial effort is taken to prevent any errors from happening in the first place. Validation is done when components are started and objects are created so that there is a minimal chance of them causing errors at later stages of operation. Likewise, checks and test routines can be written that walk through key functions and testing against a checklist of excepted responses so that the developers have an automated way of confirming that all old functionality still works when new things are added. Version numbers are one of the simplest but most important types of validation. Data formats, sub-component features and game logic can all change as upgrades take place and by versioning every aspect of a system one can track places where changes may have occurred. Likewise by storing all code and versions of all compiled components in a version control system, a history is available to track changes, allow mistakes to be undone and to make it possible to recreate exactly the state of the software at any point in its history. All these steps towards reliability are beneficial from a
scientist’s point of view. It means that any given version of the system will run the same way each time and all changes between versions can be accounted for.

5.3.2. Alien Rabble - Supporting architecture. A brief survey of some the technical elements supporting and comprising the Alien Rabble system will hopefully provide some sense of the reason why such elements were chosen and some idea of the potential that a general system built out of these elements would possess. The Alien Rabble game system is written in the Java programming language. It is a modern object-oriented language that can be run on any system. This means that, if appropriately packaged, the Alien Rabble game can be installed and run on almost any computer. There is even potential for it to be deployed over the internet. Furthermore, Java is very well supported with an excellent range of enterprise strength tools available for development, testing and version control. The code that runs the Alien Rabble was written, debugged, compiled and tested in the Eclipse Software Development Kit (www.eclipse.org), whilst version control was managed with Subversion (subversion.tigris.org). Both of these are free, open-source systems that make it exceptionally easy for multiple users to share the same code, collaborate on projects and trace the history of changes to a system. The potential of such a system should be obvious although it is also to be emphasised here that the primary objective of the project is to enable the present author to run experiments, not to turn him back into a full-time software developer or to see him become, heaven forfend, a project manager.

The Alien Rabble system itself makes great use of the fruits of a large collaborative development project that has created the jMonkeyEngine game development platform (www.jmonkeyengine.com). jMonkeyEngine was chosen for numerous compelling reasons. Firstly, the aim of the Alien Rabble project is to build as sophisticated a game engine as possible with the least amount of raw software development but whilst retaining as much control over the code as possible. A game development toolkit such as jMonkeyEngine allows a developer to leverage a large set of common and useful objects and functions that game developers need (e.g. high level rules about how cameras follow a player around, how 3D objects are lit or interact, low level things like how the pixels are written to the screen or how key strokes are handled, etc.) Furthermore, the set of tools it provides is specifically
aimed at creating immersive 3D worlds and simplified arcade style interfaces. Numerous examples and demonstration programmes showcase its capabilities and are provided to users as a starting point for their own projects. jMonkeyEngine makes use of sophisticated mathematical libraries that can give games can simulate fairly realistic physics and optics that help create a strong immersive experience. It has libraries that implement Newtonian gravity and optics and collisions with rigid and deformable bodies. It supports a rich 3D environment with ability to import objects in numerous standard modeling formats and handle sophisticated lighting and colouring effects (shadows, depth of field changes, etc.) For example, for an early demonstration version of the software it was very easy to import the well-known Greebles stimuli (Gauthier and Tarr, 1997).

Many games allow you to design your own levels and add your own features. In some cases, much of the attraction comes from the possibility of building worlds within worlds, as in the online virtual reality, Second Life (www.secondlife.com) and psychologists are already using these tools (Bailenson and Yee, 2005; Bainbridge, 2007). However, an important requirement is control over all aspects of the environment and the ability to modify the game’s functions to suit the experimenter’s needs. For example, the logging of player keystrokes to a datafile is not a typical game requirement but one that was essential to this project. This was possible with jMonkeyEngine but may not have been so with other game engines. The reason for jMonkeyEngine’s openness and modifiability is that it is a free and open source project that provides all its code for use and modification by anyone. Furthermore, it has an active online community that contributes to the continued improvement of the toolkit and provides support for the many and varied users of the system. Alien Rabble would have been simply impossible without the framework of tools, code, support, tutorials and documentation that comprise this powerful, generic game development toolkit.

5.3.3. Alien Rabble - Core set of features. The basic architecture of the Alien Rabble game allows an experimenter to import or otherwise create a set of three dimensional test objects into a various computer game-like settings. Here they can be manipulated and acted on by a child or adult playing the game while the system records the particular sequence of actions and decisions the player makes.
By providing the computer with identifiers for different categories it is possible for it to give different feedback, scores or other outcomes for different categories and thereby provide a context or situation where they can be learned. As an example, in Experiment 1 participants see an iconic smiley or sad face when they collect a 'good' or a 'bad' exemplar. The set of objects can form two or more categories. Objects can be selected and imported from a fixed library of three dimensional models (e.g. greebles), or else, a representative sampling of exemplars can be generated programmatically by building novel composite objects from some statistically defined distribution of features (as is the case with the aliens in the experiments below). The exact set of objects to be used on a given trial is controlled by the experimenter by making changes to an initialisation file.

The primary forum for interacting with the objects is a 3D virtual reality type arena. This lends an immersive quality to the game and gives opportunity for the objects to be encountered in something resembling a real world setting. Objects are 3-dimensional and can be manipulated or acted upon by the player. In the current version of the game two types of interaction with the objects are possible. In the 'grab' phase, the player uses the keyboard to drive a red grabbing device around an enclosed square arena looking for objects to collect. Any number or kind of objects can populate the arena, including test objects and distractors. In the 'sort' phase, all the objects are displayed simultaneously against a black background and a player can move them around the screen or put them into boxes, controlling the movement either with a mouse or with a computer touch screen.

Extensive user telemetry is possible. The game can timestamp significant actions and events and it can record every key press or mouse action. Data are written to XML files. This is a standard human and machine readable data format. (XML is a more generic variant of HTML). The files created can be easily read and parsed by data analysis programs (e.g., Excel, Matlab, R). This allows a substantial automatising of the often laborious and error-prone data consolidation phase of an experiment. Finally, the modular design of the game means that various features can be turned on and off as required. For example, auditory feedback and other sound effects were turned off in these versions of the game, as was a large colourful
count-down timer that could be used to put participants under time pressure. However, both of these features could be enabled by changing settings at initialisation. Likewise, other features can be added in future.

5.4. Experimental Investigation - Developmental Differences in the Learning of Rule And Similarity Category Structures

As a first test of version 1.0 of the Alien Rabble system, the following sections present two experiments that use it to look for developmental differences when children and adults encounter categories with an imposed rule or similarity structure, a third experiment compares performance in the computer game with a manual sorting task. In Experiment 1, the object is to investigate supervised learning, the game provides participants with positive and negative feedback and the aim of the game is to learn which exemplars are from which category so as to collect the ‘good’ ones and avoid the ‘bad’ ones. Experiment 2 investigates unsupervised learning, participants collect a set of items without any feedback and then are given the opportunity to sort them into between two and five groups. Performance is compared with a group performing manual sorts in Experiment 3. Prior to this we provide a brief summary of the debate about Rules and Similarity. A far more detailed account of the general debate can be found in the introductory chapter and in Section 1.17.3, therefore this section focuses more specifically on the developmental dimensions of the question.

5.4.1. Is the Rules versus Similarity distinction applicable to children’s category learning? Within adult literature, theorists generally agree that a distinction can be made between how humans learn categories based on a simple rule (e.g. all red things, objects larger than a house) and those based on a shared similarity between members of that class (e.g. dogs, chairs, Scottish accents). It is less clear to what extent these distinctions are applicable to or would be manifest in children’s categorization abilities. In this section, two theoretical positions (Ashby and Spiering, 2004; Pothos, 2005) from the adult literature are examined to see what predictions, if any, they would make about children’s performance. Following that the implications of two development accounts of category learning
In a series of papers, Ashby and colleagues (Ashby et al., 1998; Ashby and Ell, 2001; Waldron and Ashby, 2001; Ashby and Spiering, 2004; Ashby and O'Brien, 2005) present experimental, neuropsychological and neurobiological evidence for category learning being mediated by the competition between distinct brain mechanisms. Using converging evidence from development, neuropsychology, single cell and lesion studies in animals and brain scanning work, Ashby and colleagues map out a model of the brain areas involved in category learning and how they are connected and interact. Rule-like discriminations will be better handled by verbal or declarative systems and similarity judgements will be more suited to implicit or procedural-learning systems. There will also be a competition mechanism that decides which approach can solve a given problem. Ashby and Spiering (2004) map this model onto three brain areas; rule-like tasks primarily recruiting the prefrontal cortex (PFC), similarity judgments in the inferotemporal cortex (IT) and competition systems within the basal ganglia deciding which strategy is most effective in a given situation. The prefrontal cortex is one of the last brain systems to come on-line in development, and thus Ashby’s competition model predicts that young children will have difficulty in rule-switching and also in rule-like categorisation tasks. It predicts that young children might either favour similarity-based classifications that utilise non-frontal processes or else if they were to use rules they might be expected to fixate on single dimensions early in learning and perseverate despite feedback. As participants get older, the more classical distinction between rules and similarity should develop, although Ashby et al. (1998) emphasise that these two mechanisms will always be in competition.

By contrast, Pothos (2005) sees rule-like and similarity-like behaviours as opposite extremes on a continuum. Rule-like performance is associated with decisions based on the salience of just one or only a few properties of the stimulus, whereas similarity judgements integrate information across a wider number of dimensions. Whilst this account is not necessarily inconsistent with that of Ashby and colleagues, it does have a different emphasis, taking a far more abstract and theoretical approach. Pothos is interested in finding a broad conceptual framework that can be applied to rule-like and similarity-like behaviour across numerous domains. His theory has
three elements. Firstly that canonical rule-like and similarity-like processes are both extremes of the same continuum, determined by the number of dimensions that are relevant to the decision process. Secondly, that there will also be behaviours that lie in the middle of this continuum. Thirdly, similarity does not just act on perceptual inputs but can operate on more abstract properties. No developmental claims are advanced as part of Pothos’s theory but certain developmental dimensions can be implied. Since rules are said to depend upon a restricted set of dimensions, developmental differences in selective attention and response inhibition will be relevant and as these are largely mediated by developments in the pre-frontal and cingulate areas (Booth et al., 2003), the predictions may well be essentially the same as those of the Ashby et al. (1998) account. Additionally, Pothos’s third criteria that similarity can operate over abstract properties could lead to expectations that older participants are better at processing abstract relations. This is in line with Kotovsky and Gentner’s (1996) finding that between 4 and 8 years old there was an improved ability to recognise relational similarities across changes in dimension and polarity. However, none of these implications are explicitly stated in the original formulation of Pothos (2005).

As was discussed in the previous chapter (Section 1.17.3) there are two explicitly developmental accounts of early category learning abilities. Smith (1989b) proposes an holistic-to-analytic shift over the course of development while Cook and Odom’s (1992) differential-sensitivity account suggests that young children do use single dimensions but do so inconsistently and with different relative saliences. As was previously discussed, a number of studies across a range of domains appear to support the differential-sensitivity account (Ward et al., 1990; Thompson, 1994; Thompson and Markson, 1998; Schwarzer et al., 1999; Raijmakers et al., 2004) but a recent review by Sloutsky (2003) reemphasised the role of similarity. Whilst recent data from adult studies (Milton and Wills, 2004; Milton et al., 2008) make it difficult to give a straight-forward account of adult categorization abilities, suggesting that analytical responding can sometimes lead to family resemblance sorting. Finally, feedback appears to have a differential effect on rule and similarity in adults learning performance. If feedback is available but is delayed by five seconds or more, adults learning is impaired on information integration tasks but not on rule-like tasks (Ashby et al., 1999). However, if feedback is available only briefly then
learning of rule-like categories is impaired but learning similarity groupings is unaffected Maddox et al. (2004). It is not known how feedback would affect children’s learning.

Given these uncertainties, it is clear that further research into the development of rules and similarity is appropriate and that a number of questions can be asked. Firstly, since feedback during learning appears to have differential effect on adult learning of rule and similarity structures but an unknown effect on child performance, it is of interest to look at comparative performance on both category learning and spontaneous classification tasks. If there is a holistic-to-analytic shift then will children and adults show continuity in their learning of similarity-based categories and discontinuity in their learning of rule-based categories? While in free-sorting tasks, will younger children tend to favour similarity based classifications more than older children and adults? Alternatively, will the patterns of errors of the younger children support the differential-sensitivity account? For example, in learning tasks will younger children be as likely to use dimensional strategies but do so inconsistently and with more errors than older participants? Will younger participants perseverate with errors made early in learning? The next section outlines the general experimental procedures that apply to a set experiments designed to address some of these questions.

5.4.2. Experimental Investigation - General methods. This section describes the general design of the category learning game and aspects of the experimental procedures that are common to both studies presented here. All participants were tested on both experiments in a single session.

5.4.2.1. Participants. Three age groups were tested\(^1\), children in year 3 and year 6 of the British school system and a group of adult controls. Exact numbers completing each specific task are given below. Initial contact was made with the head-teacher and, once permission for testing was granted, a date was set for testing to take place in the schools. About a week prior to testing, the schools arranged for an opt-in consent form to be sent home to all the parents of children in the relevant year groups. Only those children whose parents returned these correctly filled in

\(^1\)A fourth group of children in the Reception class (ages 5-6 years) was also included in the original design, but due to poor availability, low response rates and one school failing to send out the consent forms only 8 children in this age range were tested. Their data are not included.
5.4. EXPERIMENTAL INVESTIGATION - GENERAL BACKGROUND

were tested. Children and schools were not offered any incentive to participate in the research. Although the researchers offered to give a short presentation about their research to the children. This was taken up by only one school and involved a 5 minute talk given by two researchers on the general themes in cognitive psychology together with a question and answer session with the children. After all testing was completed in a given school, the researchers wrote again to the head-teacher thanking them for their assistance and giving the school £50 of book tokens as a gesture of thanks. Adult participants were an opportunity sample collected in and around Birkbeck. Each adult participant gave their written consent before taking part in the study and was paid five pounds for their participation in both experiments at the end of the study.

5.4.2.2. Apparatus. The tasks were presented on several laptop computers (Apple Powerbook G4 15”, Apple MacBook, CompaQ NX6110) each with an external mouse connected. The game was presented in full screen mode at the full resolution of the particular computers used. However, the software scaled the scenery and objects so that they always maintained the same proportions.

5.4.2.3. Stimuli. For these experiments, the stimuli were brightly coloured three-dimensional alien beings that could vary on 9 different feature dimensions. Two examples are illustrated in Figure 5.1. Three feature dimensions (stripe colours, arm type and leg type) took on binary values that were used to distinguish two different families of stimuli used in the two different tasks. Such that a given participant would see a set of aliens from one family in the category learning task (Experiment 1) and a set from the other family in the free sorting task (either Experiment 2 or 3). The six other feature dimensions varied across a range of values. The number of legs could vary between 2 and 11. The stripe angle varied in 11 steps of 15° from -75° to 90° to the vertical. The stripe frequency could vary in unit steps from 1 to 12 cycles per standard unit. Arm length, body width and body height all varied continuously, their exact values sampled from pre-specified statistical distributions. Sets of twelve stimuli were formed using six category distributions using the two families shown in Figure 5.1. Table 5.3 lists the parameters for each of these category structures. Note that the exact sets of exemplars for a given category was fixed. There were six possible category structures, three 1-dimensional rules, a 3-dimensional rule and 2 similarity structures created by sampling from correlated
5.4. EXPERIMENTAL INVESTIGATION - GENERAL BACKGROUND

Table 5.3: The six category structures used for the aliens in Experiments 1 and 2. Within a given test set, the stripe colour, arm type and leg type were fixed whilst the other six dimensions varied to create 12 exemplars. These always divided into two 6 item categories. In each case there was one or more bimodal dimensions shown in bold with the remaining dimensions either unimodal or uniform. There were three 1D rules that were bi-modal in one number of legs, stripe frequency and body height respectively. An additional 3D rule combined these three 1D rules. Whilst two similarity categories were created by sampling from correlated bimodal distributions in arm size, body width and body height. The means of the sample distributions being either one or two standard deviations apart.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Range</th>
<th>1 dimensional Rules</th>
<th>3D Rule</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stripe Colours</td>
<td>1,2</td>
<td>Num Legs</td>
<td>Stripe Freq</td>
<td>Body Height</td>
</tr>
<tr>
<td>Arm Type</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Leg Type</td>
<td>1,2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Arm Size</td>
<td>-</td>
<td>N(1.6,0.6)</td>
<td>N(1.6,0.6)</td>
<td>N(1.6,0.6)</td>
</tr>
<tr>
<td>Body Width</td>
<td>-</td>
<td>N(8,2)</td>
<td>N(8,2)</td>
<td>N(8,2)</td>
</tr>
<tr>
<td>Body Height</td>
<td>-</td>
<td>N(13,3)</td>
<td>N(13,3)</td>
<td>N(4,2) vs N(14,3)</td>
</tr>
<tr>
<td>Leg Count</td>
<td>2 to 11</td>
<td>2,3 vs 8,9</td>
<td>uniform</td>
<td>uniform</td>
</tr>
<tr>
<td>Stripe Angle</td>
<td>-5 to +6</td>
<td>uniform</td>
<td>uniform</td>
<td>uniform</td>
</tr>
<tr>
<td>Stripe Freq</td>
<td>1 to 12</td>
<td>2,3 vs 9,10</td>
<td>uniform</td>
<td>uniform</td>
</tr>
</tbody>
</table>

Notes:
- N(\mu, \sigma) is the normal distribution
- uniform indicates a uniform dist. on the range specified
5.4. EXPERIMENTAL INVESTIGATION - GENERAL BACKGROUND

(a) Alien exemplar from family A.  
(b) Alien exemplar from family B.

Figure 5.1: Two exemplars illustrating the 9 variable dimensions of the alien stimuli. Three dimensions (stripe colour, arm type and leg type) were constant within experiment but varied between tasks. The remaining six dimensions (arm length, body height, body width, stripe angle, stripe frequency and leg count) were all varied systematically within each experiment to create the category structures.

normal distributions in three dimensions with their means either 1 or 2 standard deviations apart. An example of a single set of objects using family A and the number of legs rule can be seen in Figure 5.6(b).

5.4.2.4. Procedure. All participants took part in the category learning task (Experiment 1) and one of the sorting tasks (either Experiment 2 or 3). As many participants performed the category learning task first as performed the experiments in the reverse order. If the participant saw items from family A on the first task, they saw the contrasting set B on the second task. These were counterbalanced across participants. Moreover, if a participant saw a 1D rule in one experiment they would see a similarity-based or 3D rule in the other experiment. In both the computer-based tasks, there was an initial exploration phase where these stimuli were found in square enclosure, 150 units on the side, surrounded by a glowing "force-field" fence that prevented the participants from driving out of the game arena. Twenty-five green pyramids textured to resemble trees were randomly distributed across
the environment to act as scenery and distractors. Participants could drive directly through/under the trees. The details of how the two games differed is given in the respective procedure sections below.

5.5. Experiment 1 - Alien Grab - Category learning with feedback.

This task investigates how participants will perform when given feedback on the category membership structure imposed by the experimenter. Six different category structures were used as described in section 5.4.2.3 having either a 1D rule, a 3D rule or a similarity structure. Each participant saw 12 exemplars from a given structure with half the exemplars forming the target category. In each case positive and negative feedback was given within the game using iconic smiley faces. Participants had three rounds in which they attempted to learn which items were the good or bad exemplars. In each of the round, exactly the same items were presented. The computer recorded the order and timing with which the participants collected the objects.

5.5.1. Method.

5.5.1.1. Participants. The participants were 35 Year 3 pupils (13 female, mean age 8.4 years, SD 0.3 years ), 44 Year 6 pupils (21 female, mean age 11.4 years, SD 0.3 years ), 27 Adults (19 female, mean age 33 SD 10 years). Recruited as described above.

5.5.1.2. Stimuli. Six sets of aliens as described above. Feedback took the form of an iconic smiley face that appeared for 3 seconds when the participant captured an alien. See Figure 5.2.

5.5.1.3. Procedure. Participants were seated in front of the computer, with an experimenter seated at their side who would explain the task and monitor their performance. Initially participants were shown a screenshot similar to that in Figure 5.6(a) and it was explained to them that they would be using the arrow keys on the keyboard to drive the red grabber around the game arena. They were also shown pictures of the green smiling and red frowning iconic faces and it was explained to them that these would appear to indicate 'good' and 'bad' aliens. It was explained that their task was to collect the good aliens and avoid the bad ones. It was further
5.5. EXPERIMENT 1 - ALIEN GRAB - CATEGORY LEARNING WITH FEEDBACK. 264

Figure 5.2: Screenshot from the Alien Grab game. The player navigates the red grabber around the game arena to capture the aliens by running into them. When an alien is captured it spins, shrinks and disappears and, at the same time, the player receives feedback in the form of a green or red 'smiley' to indicate if that alien was 'good' or 'bad'. The aim is to collect all six good aliens while avoiding the six bad ones. A visual counter on the lower left indicates the progress so far.

explained that, at first, they would not know which were which but that their task was to try figure out which was which. It was also explained that they would have three attempts to do this. The experimenter checked with the participant that they had understood the instructions and then instructed them to begin the game. The experimenter observed the participant and offered prompts on how to control the movement of the grabber and to remind the younger participants of the aim of the task.

After the first round was completed the experimenter offered positive encouragement to the participant and again explained that they would now do exactly the same task again but should aim to do it both faster and more accurately. It was emphasised that the set of aliens and the good and bad rule would be exactly the same. This was repeated once more for the third and final round.
5.5.2. Results. A total of 106 participants completed Experiment 1, of these 59 took part in Experiment 1 prior to either Experiment 2 or Experiment 3 and 47 performed the tasks in the reverse order. Sixty five participants were tested using stimulus exemplars from Family A and 41 saw exemplars from Family B. In order to test the possible effects of these variables, a (2x2x3) repeated measures ANOVA was performed on the accuracy data with Experiment Order and Stimulus Set as between subject measures and Round as a within subject factor. This analysis confirmed there were no main effects of Experiment Order, F(1,102)=0.04 or Stimulus Set, F(1,102)=0.09 and neither variable interacted with Round, F’s<1. Therefore, data were collapsed across these variables for subsequent analyses.

For each participant the computer recorded the order and timings with which they collected the alien exemplars. Since each round stopped when a participant had collected all six of the ‘good’ aliens, the total aliens collected per participant per round is a measure of how well the task was learned. These results are plotted in Figure 5.2 with Sub-figures (a), (b) and (c) showing the accuracy of the participants across the three rounds, grouping them by age and grouping the response according to which type of category structure was found either similarity or 1D and 3D rules. Figures 5.4 (a) and 5.4 (b) plot the same accuracy data grouped by learning set but collapsed across all age groups.

As an initial analysis, a (3× 6× 3) repeated measures ANOVA was performed on the accuracy data grouped in this way. The between subjects variables were Age Group of the subjects at three levels (8 year olds, 11 year olds and adults) and Learning Set as a between subject variable at 6 levels (for each of the six category types) and Round as the repeated measure at 3 levels. The repeated measures data fail Mauchly’s test of sphericity, with approximate chi-squared of 9.34, p<0.002. Therefore, Greenhouse-Geisser corrected values are reported, with degrees of freedom adjusted by a factor of 0.91. There was a highly significant main effect of Round, F(1.8, 176) = 27.8, p<0.001. As can be seen from the downward sloping graphs in Figure 5.3, the participants did learn and become more accurate over the course of the three rounds. There was also a significant Round by Learning Set interaction F(9.1,176)=2.78 , p<0.005, suggesting that participants learned differently between dependent on the different category structures they were tested on. Again, this is evident from the different shaped graphs in Figures 5.3 and
5.5. EXPERIMENT 1 - ALIEN GRAB - CATEGORY LEARNING WITH FEEDBACK

(a) Mean aliens grabbed per round for similarity categories grouped by age.
(b) Mean aliens grabbed per round for 1D rule-based categories grouped by age.
(c) Mean aliens grabbed per round for the 3D rule category grouped by age.

Figure 5.3: Summary of results from Experiment 1. The graphs show the mean accuracy per round for each age group for the three different types of category structure: (a) similarity, (b) 1D rules, and (c) 3D rule.
5.5. EXPERIMENT 1 - ALIEN GRAB - CATEGORY LEARNING WITH FEEDBACK

(a) Similarity-based categories

(b) Rule-based categories.

Figure 5.4: Summary of results from Experiment 1. The graphs plot the mean number of aliens collected per round for all participants grouped by the category structure they saw. Figure (a) shows the results for the 2 similarity conditions and Figure (b) shows the four rule-based conditions.

Figure 5.5: Graph showing the mean rate of collection of aliens per round grouped by age in Experiment 1.
5.4. This is further supported by a highly significant main effect of Learning Set, $F(5,88)=3.75, p<0.004$, indicating that certain conditions were learned more easily. Finally, there was also a highly significant main effect of Age Group, $F(2,88)=9.11, p<0.001$. The older age groups were generally more accurate.

It is likely that the adults’ rapid perfect learning of the 3D rule lead to the low variance in that case. (All four adults in the 3D rule condition made no mistakes in the second and third rounds.) Therefore a further analysis excluding data from that condition was conducted. This permitted the collapsing across three rule conditions and the two similarity conditions to give a Rule/Similarity variable. A $(3 \times 2 \times 3)$ repeated measures ANOVA was conducted with Age Group and Rule/Similarity as between subject measures and Round as the within subject measure. The data still fail Mauchly’s test of sphericity, with approximate chi-squared of 13.8, $p<0.001$. Therefore, Greenhouse-Geisser corrected values are reported, with degrees of freedom adjusted by a factor of 0.86. The strong main effect of Round remained, $F(1.7, 142)=16.7, p<0.001$. But there were no significant interactions between Round and the other variables, suggesting the possibility that 1D rules and similarity are learned in similar fashion. There was a highly significant main effect of Group $F(2,82)=5.62, p<0.005$, indicating that older participants learned faster in both 1D rule and similarity conditions. However, this was mediated by a just significant Group $\times$ Rule/Similarity interaction, $F(2,82)=3.17, p<0.05$, the difference between age groups being more apparent in the rule based conditions.

Finally, an analysis was conducted on the speed with which the participants completed each round. Since the total time to complete a round would depend upon the number of items collected, the speed is better measured as at the mean time per item in a given round. Additionally, to remove the variance as participants initially learned to control the grabber, the time was taken starting from when the participants collect the first item in a given round up to the point where they collect the last good item. This was then divided by N-1 where N was the number of items collected, to give a time per item. These values are plotted in Figure 5.2 (f). A $(3 \times 2 \times 3)$ repeated measures ANOVA was conducted with Age Group and Rule/Similarity as between subject measures and Round as the within subject measure. There was highly significant main effect of Round, $F(2,164)=6.87, p<0.001$ indicating that participants were faster in later rounds. There was also a main effect of Group,
5.5. EXPERIMENT 1 - ALIEN GRAB - CATEGORY LEARNING WITH FEEDBACK.

F(2,82)=3.65, p<0.03, the 11-year-olds were generally faster than the 8-year-olds and the adults. There were no other significant effects or interactions.

5.5.3. Discussion. The results of Experiment 1 demonstrate quite clearly that the Alien Rabble category learning task is within the capabilities of the range of age groups tested, and is sensitive enough to detect differences between age groups and between different category structures. All age groups showed learning across the three rounds of the task and there were differential effects between age groups. The results showed an interesting interaction between age and rule and similarity. All age groups performed comparably when the categories were defined by a similarity structure but there was a definite effect of age when the categories were defined by a rule like structure, with older participants performing better in this latter case. This result is consistent with implications from Ashby and Spiering’s (2004) neurobiological model suggesting that younger children would have relatively greater difficulty with rule-based categories.

Children and adults do appear to learn differently. It may be that adults are solving the rule-based tasks with an explicit hypothesis testing approach whilst the children operate more implicitly. The perfect performance of all the adults in the second and third rounds of 3D rule and their high degree of accuracy on the 1D rules suggests that they rapidly are acquiring the correct rule in a way that the children are not. However, this could just be a matter of degree not a qualitative difference. Since the adults appeared to learn rapidly in the rules based conditions, their accuracy was close to ceiling and therefore the current experiment cannot distinguish if their superior ability is the result of a different learning strategy. The advantage of the computer game experimental design is that we can potentially adjust this factor to investigate the question, either by making the difference between the dimensions smaller, increasing the number of dimensions that vary or possibly introducing other features to make the task harder for all participants.

Alternatively, it could be asked why the children appear not to be using rules. Is it because they are processing the stimuli more holistically (Smith, 1989b) or are they less sensitive to the particular dimensions being tested (Cook and Odom, 1992)? The results of Thompson (1994) and Raijmakers et al. (2004) suggest that children can use rules but do so inconsistently. The results of the free sorting
5.6. EXPERIMENT 2 - ALIEN SORT - FREE CATEGORY CONSTRUCTION

task in Experiment 2 (described below) could potentially provide information on how sensitive children are to the various dimensions of these particular stimuli. Alternatively, it might be possible to gain some information from an analysis of the item by item choices that children make during the course of the task. This could be achieved by looking for evidence that they perseverate on a particular contrast or feature when collecting the objects (e.g. going repeatedly for tall exemplars). Or else it might be possible to find evidence that the younger participants cannot inhibit a prepotent tendency to go for the next nearest item irrespective of its features. Both of these possibilities can potentially be investigated using the data available from the task. However, both analyses would be likely to require larger sample sizes than in the present study to be meaningful. Moreover, at present, it is unclear exactly how such analyses might be conducted. This is left as a question for future research.

One final point of interest was the finding that 11-year-olds were the fastest at performing the task, whilst the 8-year-olds were no slower than the adults. This supports the original hope that computer games can provide an appropriate means of testing young children. The participants in this task rapidly understood how to control the game and the task objective.

5.6. Experiment 2 - Alien Sort - Free category construction with prior familiarisation

The second experiment aims to investigate the kinds of groupings participants would spontaneously apply to these sets of stimuli and whether there would be any developmental difference in the types of groupings they produced. As in Experiment 1, three age groups were compared, 8-year-olds, 11-year-olds and adults. Since the stimuli varied on 6 dimensions, each of which varied in a scalar across a range of values, the range of potential classifications is much greater than in a typical categorization study with artificial stimuli, such as Medin et al. (1987). Therefore, participants are permitted to divide the stimuli into a larger number of groups (between 2 and 5). Additionally, to give the participants some interaction with the stimuli before classification and to emphasise the game-like nature of the task, there was an initial collection phase prior to classification.
5.6.1. Method.

5.6.1.1. Participants. The participants were 29 Year 3 pupils (mean age 8.4, SD 0.3 years), 29 Year 6 pupils (mean age 11.3, SD 0.3 years) and 24 Adults (mean age 33.6, SD 11 years). All participants also completed Experiment 1. The order of presentation was counterbalanced by random assignment.

5.6.1.2. Stimuli. The same six category structures were used as Experiment 1 (see Table 5.3). However, if participants saw sets from a different family than they had seen in Experiment 1. Likewise, if participants had seen a rule-based category type in the first experiment they saw a similarity-based category in this experiment and vice versa.

5.6.1.3. Procedure. Participants were seated in front of the computer, with an experimenter seated at their side who would explain the task and monitor their performance. Initially participants were shown a screenshot similar to that in Figure 5.6(a) and it was explained to them that they would be using the arrow keys on the keyboard to drive the red grabber around the game arena. Their task was to collect all the aliens as quickly as possible by driving into them. The experimenter observed the participant and offered prompts on how to control the movement of the grabber. When each alien was collected it would spin round once whilst shrinking and then disappear, a counter in the lower left indicated how many aliens remained.

Upon completion of this first familiarisation phase, the experimenter advanced the game into the sorting phase. The twelve aliens collected in the initial phase were displayed in a horizontal line at the top of the screen in a random order and there were 3 coloured boxes along the bottom of the screen. See Figure 5.6(a). The experiment explained that the task was to sort the aliens into groups by putting them into boxes. The participants were shown how to do this by clicking on an alien to select it and then clicking on the box they wished to put it in. They were told that there were no right or wrong answers but that the aliens should be put into smaller groups that looked liked they went together. They were told that they could have between 2 and 5 groups and there could be any number of aliens in a single box and that they could take as long as they wanted. At this point the experimenter also showed the participant how to add and remove boxes using the up and down arrow keys. The experimenter checked that the participant understood the instructions
5.6. EXPERIMENT 2 - ALIEN SORT - FREE CATEGORY CONSTRUCTION

(a) Alien collection phase. The player navigates the red grabber around the game arena to capture the aliens by running into them. When an alien is captured it spins, shrinks and disappears. A counter on the lower left indicates how many aliens remain.

(b) Alien sorting phase. The 12 aliens collected in the previous phase are displayed in random order in a line in the upper part of the screen. The player must sort them into families. Using the mouse, the player clicks on an individual alien which moves to the centre of the screen and slowly spins so the player can examine it. It can then be sorted by click on one of the boxes in the lower part of the screen to place the alien in that box. Boxes can be added and removed, allowing between two and five groups in total.

Figure 5.6: Screenshots from the Alien Grab game used in Experiment 2.
and then remained nearby to monitor their performance and provide assistance if necessary. Upon completion the participants were thanked for their participation. The computer recorded the participants choices to a data file.

5.6.2. Results.

5.6.2.1. Determining the sorting criteria in a free sorting experiment. In the current experiment, the participants were allowed more freedom to form multiple groups than typically found in a standard classification task. Likewise the dimensions of the objects varied in a scalar fashion rather than simply taking binary values. Therefore, the resulting groupings will need to be analysed in a more complex fashion than in previous studies (e.g. Medin and Schwanenflugel, 1981; Regehr and Brooks, 1995). Moreover, any such analysis ought to be based on some objective quantification of the category structure. To achieve this a new measure of dimensional predictiveness is introduced, based on the idea that a good category
structure will minimise the within category variation and maximise the between category variation (Rosch, 1975).

If a participant is using variation on a given dimension to sort the items then that dimension should have a large between category variance score and low within category variance score. By contrast if a particular dimension is ignored by a participant then on average that will have a moderate between category variance and a moderate within category variance. Thus, a predictiveness score for each dimension can be derived by taking the difference between these two values (Equation 5.17). This value will be high when the dimension is predictive within the classification and low otherwise. An example for two possible dimensions is illustrated graphically in Figure 5.8.

\[
(5.17) \quad \text{Predictiveness} = \text{Variance}_{\text{Between}} - \text{Variance}_{\text{Within}}
\]

If the range of variation on each dimension is scaled to a standard range of between 0 and 1, it should be possible to compare the relative predictiveness of all of the dimensions. If \( G \) is an individual’s grouping of the stimuli and \( g \) and \( h \) denote subgroups within that sorting, with \( x_{gi} \) being the individual values for each element in \( g \) and where \( \bar{x}_g \) is the mean within group \( g \). Then Equation 5.18 gives the formula for the total variance within the sub-groupings.

\[
(5.18) \quad \text{Variance}_{\text{Within}} = \sum_{g \in G} \left( \sum_{i \in g} (x_{gi} - \bar{x}_g)^2 \right)
\]

Additionally, if \( |G| \) denotes the size of the set of all stimuli, \( G \) and is the \( |h| \) is the size of an individual subgroup \( h \). Then Equation 5.19 gives the formula for the total variance between the sub-groupings. The term involving \( |G| \) and \( |h| \) is a scaling factor introduced so that both variance scores are relative to the size of the group \( G \), irrespective of how many sub-groups there are.
Table 5.4: Table showing distribution of the most predictive dimension for each participant across category structure conditions. The final column indicates the total number of participants in each condition.

<table>
<thead>
<tr>
<th>Rule - Leg Count</th>
<th>Arm Size</th>
<th>Body Width</th>
<th>Body Height</th>
<th>Leg Count</th>
<th>Stripe Angle</th>
<th>Stripe freq</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule - Stripe Freq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule - Body Height</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule - Leg Count -</td>
<td>Track 3</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Stripe Freq &amp; Body</td>
<td>Stripe Freq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarity - 1st Arm Size, Body Width, Body Height</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similarity - 2nd Arm Size, Body Width, Body Height</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

(5.19) \[ \text{Variance}_{\text{Between}} = \sum_{h \in G} \frac{|h|}{|G| - |h|} \left( \sum_{i \in g, g \neq h} (x_{gi} - \bar{x}_h)^2 \right) \]

5.6.2.2. Sorting results from within the game. First, it is observed that all age groups used comparable number of groupings. On average, the Year 3 pupils divided the stimuli into 3.7 groups, Year 6 used an average of 3.7 groups and the adults used an average of 3.6 groups. Next, the dimensional predictiveness scores for each participant were calculated, using the formulas described in the previous section. Figure 5.8 shows these data grouped by category type\(^2\). As can be seen from the graphs, the leg count and stripe frequency dimensions tended to dominate classifications when either one, the other or both of these dimensions were bimodally present in the category structure, (Sub-figures (a), (b) and (d) respectively). In the condition where the category structure was bimodal in body height, this dimension was moderately predictive and to a lesser extent groupings also reflected variation in leg count and stripe frequency as seen in Figure 5.8 (c). While for the two similarity based classifications (Figures 5.8 (e) and (f) ), no clear pattern emerges, although stripe frequency appears to be a salient dimension in the 1 standard deviation similarity case. . Table 5.4 represents the same data in a simplified way; It makes a count of the single most predictive dimension per participant grouped according the category structure they experience.

\(^2\)Because of the small group sizes there were not enough data points to compare across age groups across conditions.
5.6. EXPERIMENT 2 - ALIEN SORT - FREE CATEGORY CONSTRUCTION

(a) 1D rule - Leg Count

(b) 1D rule - Stripe Frequency

(c) 1D rule - Body height

(d) 3D Rule - Leg Count + Stripe Freq + Body Height

(e) Similarity structure - 1 std. dev. bimodal in arm length, body height and body width

(f) Similarity structure - 2 std. dev. bimodal in arm length, body height and body width

Figure 5.8: Mean dimensional predictiveness scores for the six category structures sorted by all game participants in Experiment 1B. Error bars are 95% Confidence intervals.
Using the dimensional predictiveness scores, it is also possible to look at whether the sortings that the participants make are uni-dimensional or multi-dimensional. If the first and second most predictive dimensions has a comparable scores then it is likely that both are being sorted for. The level of comparability is determined by taking a cut-off. Thus, if a cut-off of 50% is used, if no other predictiveness score is greater than 50% of the largest predictiveness score then it may be said that the classification was uni-dimensional. Since the cut-off chosen is somewhat arbitrary, Table 5.5 shows these values calculated with three different levels of cut-off. It can be seen from the table that in all cases, the distribution across age groups is approximately the same. To confirm this independence of classification type and age, separate chi-squared analyses were conducted at each of the levels of cut-off, comparing the number of uni-dimensional versus multi-dimensional sortings for each age group. For the 50% cut-off, $\chi^2 = 0.75$, d.f. = 2, p < 0.70. For the 66% cut-off, $\chi^2 = 0.69$, d.f. = 2, p < 0.71. For the 75% cut-off, $\chi^2 = 0.57$, d.f. = 2, p < 0.75.

5.6.3. Discussion. This experiment investigated how individuals would freely sort a set of 12 stimuli which varied on 6 dimensions with one of six possible category structures (see Table 5.3) which they encountered in a computer game setting. Participants used the computer interface to sort the exemplars into between 2 and 5 groups. Because of the potential complexity of the groupings the participants could produce there was no simple means for an experimenter to classify participants’ sorting strategies. Therefore, a novel measure of dimensional predictiveness was introduced, based on Rosch’s (1975) idea that categories maximise between group difference whilst minimising within group variation.

Table 5.5: The distribution of uni-dimensional and multidimensional sorts shown for various levels of cut-off the relative importance of the second dimension. See text for fuller explanation.
Using this measure, it was found that in some cases participants can discover some bi-modal category structures in unsupervised task. For the 3D rule-based category structure and the 1D leg count and stripe frequency rules the dimensional predictiveness measure clearly indicated that participants were sorting primarily on the basis of these dimensions. There was less evidence of a favoured dimension for sorting in the three other conditions (1D rule on body height, 1SD and 2SD bimodal similarity structure on arm length, body width and body height). Additionally, by comparing the relative values of the most predictive and the next most predictive dimension provided a criterion for determining if classifications were unidimensional or multidimensional. Since the cut-off level of this difference was arbitrary, the data were analysed with cut-off levels of 50%, 66% and 75% between the dimensions. It was found that at all levels of cut-off, all age groups make comparable use of uni- and multi-dimensional sorting strategies. This is at odds with the holistic-to-analytic-shift account (Smith, 1989b).

To some extent, all analyses were hampered by small number of participants. There were only around four participants per age group per condition and sometimes even fewer participants in particular cases (see Table 5.5). This made it impossible to compare the performance across ages on the individual category conditions in a statistically rigorous fashion. This problem was anticipated at the time of testing but due to the relatively low response rates to the opt-in consent and the timing of the school holidays meant that it was not possible to recruit any more school-age participants in the time available. However, it should be noted that by allowing multiple groups in the sorting task and by the using dimensional predictiveness measure the experiment provided a richer dataset than the categorical data normally collected during classification tasks.

Furthermore, multiple groups permit the classification to capture hierarchical structure. Participants can make gross divisions on the basis of highly salient features (e.g. tall versus short) and then subdivide these groups further according to secondary considerations (e.g. broad or narrow stripes) or else hierarchical classifications may be even more complex (e.g. height, etc. may be overlooked for an anomalous subgroup with vertical stripes). In a classification task where participants are allowed to sort into just two groups, there may well be a demand characteristic to find a single feature and ignore other sources of potential grouping information.
Furthermore, note that the 3D rule case, the variation on body height, leg count and stripe frequency are perfectly correlated to a first order. Namely, any spilt of the items into exactly two groups according to the bimodal distribution of any one dimension of out of body height, leg count and stripe frequency would also partition the group by the other two. However, if participants are allowed to choose multiple groups then second order differences between these dimensions can be captured.

This suggests that the dimensional predictiveness measure of between category variance minus within category variance is a sensitive measure of classification. However, it falls short of capturing everything that is known about categorization. Firstly, it is known from work of Shepard (1957; 1987) that the psychological similarity of two stimuli is a monotonically decreasing function of the distance function that describes their separation. The current measure does not capture this. Secondly, it treats dimension independently and in parallel. Therefore, it is a prototype model and although it detects the individual predictiveness of correlated dimensions (as in the case of the 3D rule), it will fail to give weight to the correlational structure. This is an important feature captured by exemplar models such as the Context Model (Medin and Schaffer, 1978) and its successor the Generalised Context Model (GCM - Nosofsky, 1986). Thirdly, unlike in GCM and similar models, the dimensions are not weighted. However, this is parsimonious in that it makes no assumptions about the weightings for each dimension and these will be potentially be different between participants, making it difficult to fit any function with free weight parameters. Moreover, accurate weighting parameters are less important when dimensions are treated independently and because each parameter has been scaled to range between 0 and 1, the relative variance and predictiveness can actually give a measure of dimensional salience. As appeared to be the case the Stripe Frequency was a highly salient dimension for participants in this study, whilst Stripe Angle was largely ignored. In general psychological dimensions experienced need not be identical to ostensible parameters of the objects and the fitting of dimensional weighting depends upon the choice of both distance metric and the similarity function that are used (Gärdenfors, 2000). However, this discussion is beyond the scope of the present investigation.
5.7. Experiment 3 - Manual free-sorting task

In addition, it was also of interest whether use of the game environment would by itself induce any different sorting to a manual card sorting task. At worst, the game environment may add extra cognitive load and distraction from the task, preventing the participants from classifying the exemplars as accurately as they were capable of doing. This would particularly be a concern when testing younger age groups. At best, the game environment may in and of itself be sufficiently enriched and engaging that participants sort exemplars in a totally different way to more traditional laboratory tasks. Obviously, these possibilities can only be fully addressed with an extensive task analysis and validation along the lines suggested by Markman and Ross (2003). But as a initial exploration, a further group of 11-year-olds took part in a manual card based version of the sorting procedure.

5.7.1. Method.

5.7.1.1. Participants. The participants were 24 x Year 6 (mean age 11.4 years, SD 0.25 years) who also took part in Experiment 1.

5.7.1.2. Stimuli. For the manual sorting task, 6 sets of 12 laminated cards were made up, one set for each of the six category types described in table 5.3, using aliens from Family B. All aliens were presented in full colour on a plain mid-grey background. The cards were rectangular with a 1cm margin beyond the maximum extent of the aliens, which varied in total height from 7.2cm to 13.7cm and in total width from 4.9cm to 17.7cm.

5.7.1.3. Procedure. Children were tested in groups of 2 or 3 either before or after taking part in Experiment 1. Each child was seated at a separate desk with arranged so they could not see each other’s cards. Each child was presented with the 12 cards from the appropriate category type laid out in a random order in front of them. They were told to sort the cards into smaller groups that looked like they went together. They were told there were no right or wrong answers, that they could have any number of groups of any size\(^3\) and that there was no time

\(^3\)In Experiment 2, the computer allowed participants to sort the items into a maximum of 5 separate groups. However, it was decided that imposing a limit of 5 groups in the manual sorting task would make the task instructions overly complicated because it required the participants to hold an arbitrary maximum limit in memory. As it turned out there was only one participant who used more than 5 groups (that participant used 6 groups).
5.7.2. Results.

5.7.2.1. Comparing in-game sorting to manual sorting. Participants in Experiment 3 sorted the stimuli into an average of 3.5 groups. This was comparable with the 3.7 groups produced by the Year 6 pupils in Experiment 2. To compare the performance between these groups, the same predictiveness scores were calculated and the averages and confidence intervals per category type for each of the two groups were plotted. Figure 5.9 shows these results for the three category structures for which sufficient data was available.

Table 5.6 compares the number uni- and multi-dimensional sorts produced in-game and in manual sorting task. As before these values were calculated with three different levels of cut-off. Chi-squared analyses were conducted at each of the levels of cut-off, comparing the number of uni-dimensional versus multi-dimensional sortings for each age group. For the 50% cut-off, $\chi^2 = 2.14$, d.f.=1, p<0.14. For the 66% cut-off, $\chi^2 = 0.62$, d.f.=1, p<0.43. For the 75% cut-off, $\chi^2 = 3.17$, d.f.=1, p<0.07. This last comparison was marginally significant but this maybe due the lack of participants in the 1D Stripe Freq. rule condition in the manual sort task.

5.7.3. Discussion. The small number of participants makes it difficult to draw any strong conclusions. Nevertheless, it seems clear that there was a very similar pattern of sorting between a screen-based sorting task and a card-based task when using the rich stimuli. This is encouraging because at this stage the limit to decide. Upon completion the experimenter thanked the participant for their participation and then manually recorded the groupings that each child had created.
5.7. EXPERIMENT 3 - MANUAL FREE-SORTING TASK

(a) 1D rule - Leg Count

(b) 3D Rule - Body Height, Leg Count + Stripe Freq

(c) Similarity structure - 2 std. dev. bimodal in arm length, body height and body width

Figure 5.9: Comparing the mean dimensional predictiveness scores in game (light blue) and manual (red) sorting tasks by Year 6 participants. There were insufficient data for the other comparisons. Error bars are 95% Confidence intervals.
computer-based sorting task is literally just a direct analogy of the card-based task. In particular, we have not yet attempted to further disguise the nature of the task by building more of a story into the game or added objectives over and above the classification of the set of exemplars, or make other manipulations that may improve the ecological validity of the computer based tasks. As Markman and Ross (2003) warned, it is necessary to know how any new tasks and techniques for investigating category learning relate to the existing literature both theoretically and empirically. This control task is a first step in that validation process.

5.8. General Discussion

In Experiment 1, 8 year old, 11 year old and adult participants were required to collect positive and avoid negative exemplars in a category learning task. Participants were tested across 6 different conditions that presented categories based on either a 1-dimensional rule, a 3-dimensional rule or a similarity structure. It was found that participants could learn in all cases except where categories were based on bimodal similarity structure one standard deviation apart. Performance on similarity based tasks was comparable across age groups but adults were markedly more accurate in the rule learning conditions. However, in terms of speed of completing the task 11-year-olds were the fastest group while adults were no faster on average than 8-year-olds. It should be emphasised that participants at all ages showed evidence of learning. Thus, this computer-based technique provided a direct test of category learning in young age groups and importantly one not based on the triadic classification task. Furthermore, it is notable that one of the very few other developmental studies to use a classification task with early school age children (Livingston and Andrews, 2005) also used 'friendly-looking aliens’ (p.320, ibid) and embedded the task in an ecologically rich context.

In Experiment 2, the same groups of participants used a computer interface to sort a new set of aliens into between 2 and 5 groups of aliens that ‘look like they go together’. Unknown to the participants, the 12 exemplars once again came from one of six bimodal category structures. The participants’ classifications were analysed using a newly developed measure of dimensional-predictiveness. Analysis showed that in some cases participants do detect and classify according to the category structure imposed by the experimenter, and this was more likely in the
rule-based categories. Furthermore, the distribution of uni-dimensional and multi-
dimensional sorts was similar across all three age groups. A control group of 11-year-
olds manually sorted printed cards featuring the same sets of aliens (Experiment 3). 
Although the small numbers involved prevented a detailed statistical comparison,
this group showed a similar set of classification patterns.

The most general conclusion that can be drawn is that the games 'work'. They
answer the original objective of developing novel methods that are suitable for
investigating the development of category learning in young school age children.
Children as young as 6 years old were tested using the game-based tasks. They
understood the tasks. They could operate the controls and above all seem to
enjoy and engage with the games. In fact, participants of all ages enjoyed and
engaged with the tasks and the games were particularly attractive to modern highly
computer-literate children. The children were no slower and often faster than the
adults and their comments confirmed their enthusiasm for the task; e.g. “I know
this game!”, “Can I get this on my computer?”, “Are we doing this every Tuesday?”

Furthermore, the games provided data showing both similarities and differences be-
tween adult and child participants' learning of rules and similarity based categories.
Of particular interest are the apparently conflicting results of the two experiments.
In Experiment 1 adults are significantly better at learning a rule based category
than the children but are comparable with similarity structures. Whereas, in Ex-
periment 2, no effects of age were found; adults and children used a comparable
number of subgroups to classify the stimuli and produced similar distributions of
uni- and multi-dimensional classifications. It is not clear how either Smith’s (1989)
holistic-to-analytic-shift theory or Cook and Odom’s (1992) differential sensitivity
theory would fully account for this pattern of results.

Looking in greater detail at the difference may clarify this. Results from Experi-
ment 2 suggest that younger children can discriminate and classify uni-dimensional
rule structures as well as adults. The difference appears to be that adult rule learn-
ing shows an abrupt rapid learning while adult similarity and all child learning
is more gradual. This leads to a question of whether the qualitatively different
profile of adults superior rule-learning performance was due to a different learning
mechanism or if alternatively it was a consequence of a ceiling effect. In either
case, an explanation must be given for why this difference occurs with rule based not not similarity based category structures. It could be that adults are exhibiting fast, insightful *rational* learning via a process of hypothesis testing whereas children learn more gradually. Schmittmann et al. (2006) assessed the performance of young school-age children and undergraduates in the Wisconsin Card Sorting Task (WCST, Heaton et al., 1993) using Finite Mixture Models and found evidence for two distinct learning processes with abrupt learning more likely in older participants. It may be possible to perform an equivalent analysis on a redesigned version of the current task. However, this approach is not without difficulties, as the current task provides a far wider range of potential hypotheses for a rational model to consider. In any case, it would be instructive to rerun the task with harder rule based discriminations to investigate ceiling effects.

This would require a clearer idea of the effective psychological dimensions of the stimulus sets, which, in turn, leads into a consideration of what those psychological dimensions are and how they might be determined. The results and analyses of Experiment 2 show that the use of the dimensional predictiveness measure can provide a means of quantifying the dimensions that are relevant to a participant’s classification. However, as was discussed in Section 5.6.3, this approach does not capture the full complexity of category structure. A review by Goldstone (1994) emphasises that while judgments of similarity are not fully explanatory for categorisation behaviour, they can provide a principled grounding for it. Thus, a complement to future research using the tasks described above would be to find an equally engaging game scenario that would allow the collection of similarity ratings under a range of conditions. (One possibility might be a matching-to-sample game where proportionally more points are awarded for finding better matches from a range of alternatives). Quantitative similarity data would open up the possibility of fitting dimensional weightings and similarity functions. This would not be straightforward. One of the main motivations of Nosofsky’s original GCM paper was that “perceived similarity (or distance) is *not* invariant across different paradigms” (p. 56, Nosofsky, 1986) and parameter fitting in that paper was achieved by means multi-dimensional scaling of very large amounts of data. It is anticipated that with access to this quantitative similarity data, the dimensional predictiveness approach could be superseded by a more complex exemplar model along the lines of GCM.
If this can be achieved in a developmental context, Alien Rabble would become a very powerful investigative tool.

5.9. Future Directions

Given the original ambitions of this chapter and the successes with the studies described here, it is natural to consider the possible future directions that the development of the Alien Rabble approach should take. In this section various possibilities are considered, some more speculative than others. Both Markman and Ross (2003) and Murphy (2003) emphasise that while new approaches are needed in the study of category learning, it is important to relate new work to existing models and paradigms. Therefore, one crucial theme for the evolution of this tool will be to assess its validity and usefulness as a new experimental tool and to carefully evaluate the results it produces. While the work presented here has been largely exploratory, it is hoped that it already points towards a rigorous methodology; In cases where they are appropriate traditional card-based control tasks are to be used and there is a clear commitment to a quantitative approach to category structure.

Indeed, the computer-based nature of the framework makes it very well suited to a highly quantitative approach. The system allows the construction of sets of objects with multiple independently varying scalar and feature-based dimensions the parameters of which can be manipulated online. This would permit very wide sampling from the object space and the creation of artificial category structures based on correlational considerations, on family-resemblance variation from a prototype or various other parametrisations. It is potentially even possible to scale variation along a particular dimension based on individual’s psychological similarity functions or selectively generate exemplars that probe the evolution of participants perception of a category boundary. This highly quantitative approach to stimuli is complemented by equivalent approach to responses. The system can provide very extensive telemetry of a participants interaction with the computer. This rich data collection could allow wider range of phenomena to be investigated. For example, in sorting tasks it would be possible to look at a participants step-by-step choices investigate prototypicality and order effects. Do participants initially favour prototypes when determining categories? Do they form contrast groups, initially picking
the most dissimilar items and organizing around those? As another example, in Experiment 1, it was observed that, at various points, many of the participants would approach and inspect multiple exemplars before making a particular choice. This may be indicative of a hypothesis testing approach. With more extensive logging of the state of the game at these intermediate points, it may be possible to investigate how such processes unfolds over time and learning.

Indeed, one of the main aims for future work in this area will be to pay greater attention to the process of learning. It is hypothesised that developmental differences in learning mechanisms could be one of the main explanatory factors behind developmental differences in rule- and similarity-based categorisation. In Millward and Wickens's (1974) model of Concept Identification learning two types of learning were possible; a gradual associational mechanism and a hypothesis-testing mechanism. Raijmakers et al. (2001) found evidence that longitudinal discrimination learning data from 6-10 year old children was best fitted by a mixture model with rational and gradual learners. Whilst Schmittmann et al. (2006) reached a similar conclusion about the WCST. The current studies could be adapted to make them amenable to similar finite mixture modelling techniques. Such approaches require larger samples of several hundred participants. The automatisation of the game task and data collection together with potential to be easily deployed to every computer in school computer lab or accessed over the internet make the collection of large sample feasible and efficient.

On a related theme, the intention is to make the software and source code available to other labs. Already the system has been used at the University of Cardiff in a final year undergraduate project (Salt, 2008). That study assessed how assigning a functional role affected sorting behaviour of undergraduate participants. It used different materials which had the category structure from Medin et al. (1987) as shown in Table 5.1. Half the participants were given simple instructions similar to Experiment 2 while the other half were given a more elaborate scenario that they were scientists who had to collect and sort aliens by family. A further two groups performed equivalent manual card-based versions of the task. The experiment found an interaction between pre-exposure and role assignment that lead the participants to make more complex sorts, suggesting these factors have an important effect on ecological validity. The software also met with positive response when presented
at a recent British Psychological Society, Developmental Conference (Addyman, 2008).

Conversely, it may be appropriate to use the system with stimuli and/or tasks provided by other labs. For example, the system is capable of recreating the programatically generated sea-shells that Gärdenfors (2000) used to investigate various similarity metrics. Or it may be possible to use this system to build bridges between results that look at expert categorization or object recognition. The Alien Rabble system allows the import of objects in many standard 3D modelling formats and these can then be scaled, rotated and otherwise manipulated within the game environment. Thus it is entirely feasible to conduct studies using the Greebles, YUFO’s and other artificial categories of objects that have been used in the extensive experimental work investigating perceptual expertise (Gauthier and Tarr, 1997; Rossion et al., 2004; Vuong and Tarr, 2006). If the difference between the laboratory and the real world is of crucial importance then even the current study can be criticised for using artificial and low-dimensional objects. One possibility might be to utilise the Dryad package (website: dryad.stanford.edu) developed by the Stanford Virtual Worlds research team (Talton et al., 2008). This is a tool that generates highly realistic three dimensional models of trees based entirely on a set of nearly 100 numerical parameters. These form the latent dimensions of the category space at least as complex as that investigated in Medin et al. (1997).

One final extremely important development will be to ensure the best possible accessibility for very young children. Due to administrative, timing and availability issues it was not possible to test a large enough group of Reception class children (ages 5-6 years) for inclusion in the present studies. However, around 8 children in this age group were tested on each of the experiments. As pilot work, the results from these participants were very encouraging. Almost all children were able to complete both tasks. However, it was found that this age group did need considerably more prompting on how to use the keyboard and reminding of what the task required. This could introduce some variation between participants. As Gee (2003) describes, it is a common feature of modern computer games to lead the player through the necessary learning steps and adjust the task to meet their current level of competence. It is hoped that versions of Alien Rabble will adopt this pedagogic approach.
This chapter presented the final set of experiments within this thesis. In the following chapter, the foregoing experimental work is summarized and some general conclusions drawn in light of the questions originally posed in the introduction.
CHAPTER 6

Conclusions

Four strands of research were conducted addressing the topic of Rules and Similarity in category learning from a number of different perspectives. Three sets of experiments looked at the emergence of various abilities in infancy. In the final experiment a pair of novel computer games was specially developed to investigate the use of the rules and similarity by adults and school-age children. In this chapter, the main findings of those experiments are briefly summarised and a synthesis is attempted that relates those specific findings back to the more general questions and themes laid out in the introduction. From here a set of questions for future research are raised. The following section reflects back on the wider lessons learnt in the course of planning, carrying out and writing up this research. The final section draws this thesis to a close with a few discursive remarks and speculations.

6.1. Summary of Findings

6.1.1. Visual Sequence Learning. Artificial grammar learning (AGL) was chosen as the first topic for research because of its use investigating the early precursors of linguistic abilities and other simple rule use. The research built on a visual AGL paradigm originally devised by Kirkham et al. (2002). Four experiments were conducted: The first three experiments investigated five month old infants’ spontaneous preferences when presented with two visual sequential grammars of differing complexity whilst Experiment 4 required adults to judge which of two paired sequences was more random.

The infant studies found that 5-month-olds spontaneously looked longer at more complex, less orderly grammatical sequences. An item by item analysis of the infants’ attention to the screen supported the idea that they were not learning global properties of the grammars presented, but instead responded to local redundancy and possibly also small fragments of the grammar. Infants were more likely to look
away from the screen when there was a short repetitive sequence or when they encountered items/pairs of items that had occurred before with a high frequency. There was an unexpected main effect of gender with male infants looking longer overall in Experiment 1 and a trend in the same direction in Experiment 2 but this effect went away in Experiment 3. There are no clear theoretical accounts of such a difference. The adults in Experiment 4 showed the same discrimination failure and successes as the infants in Experiment 1 and 3. They were shown to be likely to be using explicit, often verbal, strategies based on local features of the test sequences.

The study showed that even on simple tasks infants’ behaviour can be complex. This was revealed by detailed moment by moment analysis of data. A review of adult AGL literature by Pothos (2007) identifies two important orthogonal dimensions that ought to be considered when assessing AGL research; the implicit-explicit representation of knowledge and rule use as defined by frequency independence. The present study found no evidence of rule learning and in fact found that local statistical features had a greater effect on performance than global frequency. This local effect points to the importance of working memory in the infant performance, suggesting that infants are explicitly representing the sequences. Further support for this interpretation comes from the analogous pattern of discriminations found in the adult task.

6.1.2. Abstract Same/Different Concept Learning. The next chapter looked at abstract Same/Different (S/D) learning in young infants. This provides one of the simplest examples of an abstract rule and is a pre-requisite in the rule-based AGL tasks such as that presented by Marcus et al. (1999). Furthermore, the two item Same/Different (S/D) relation had previously been extensively investigated in the animal cognition literature.

Two new experimental paradigms were developed; a passive habituation task with visually rich photographic stimuli and an active anticipatory eye movement (AEM) task that uses simple monochrome geometric shapes. Two groups of infants, aged 4 and 8 months, were tested to provide a developmental perspective. In the habituation task, groups of infants were habituated to either Same or Different stimuli pairs. At test, all individual images were novel and 8-month-olds but not 4-month-olds dishabituated when the test stimuli showed a novel S/D relationship. In the AEM
task, 4- and 8-month-olds both responded in a similar fashion. They rapidly learned to anticipate the re-emergence of occluded Different paired shaped but performed at chance when the pair of shapes were the same as each other. Additionally, both age groups transferred their learning in the Different case in generalization trials with novel coloured shapes. This pattern of responding is not consistent with a full awareness of the S/D relation but may be in response to some lower level other feature of the stimuli.

The pattern of results was surprising. On one hand, the habituation studies provided strong evidence that by eight months old infants can use an abstract S/D relation which generalises across an otherwise completely unrelated set of stimuli. On the other hand, the AEM studies suggested that both 4 and 8 month old infants are responding to some lower level perceptual features. The different experimental set-ups may account for this. In particular, it might be important that habituation is a passive task whilst AEM requires active searching. These results provide continuity with results in the animal literature and, contrary to a primarily conceptual interpretation (Mandler, 2004), it appears that infants’ S/D concepts are firmly grounded in perception.

### 6.1.3. Family-resemblance or unidimensional category learning

In chapter 4 three experiments attempted to see how infants respond to rules and similarity. Infants were habituated to sets of stimuli that contained both unidimensional (rule) and family resemblance (similarity) category structures, and their reactions were measured to a set of conflict trials that placed these two sources of information in opposition to each other.

In Experiment 1, 11-month-olds were presented with artificial stimuli composed of coloured geometric shapes, each of which resembled a schematic house. All the stimuli had a consistent single feature (an internal white arrow that pointed to the right). The infants habituated over the initial trials. In the test phase, they did not dishabituate to either the family resemblance or single feature test item but did dishabituate to a completely novel bunny rabbit. This suggests that all habituation and test items were treated as members of one category. The stimuli were all simple compounds made of similar components and may have been that the test manipulations were too subtle for the infants or the stimuli too boring.
In Experiments 2A and 2B, 11- and 5-month-olds were habituated to photographic stimuli all from a single global category (all houses or all cars) and each of which had been digitally manipulated to feature the same highly distinctive texture (as either roof or windscreen). The test stimuli showed either an item from a novel family with the original texture or a new item from the original family with novel texture. There was an interesting asymmetry, in that responses depended on the nature of the habituation set. Infants familiarised with houses dishabituated to both family resemblance (novel category) and unidimensional (novel texture) test trials, whilst infants familiarised with cars did habituate and only looked longer at the completely novel test trial. This pattern was found consistently across both the age groups suggesting that this was a real effect caused by the difference between the familiarisation sets.

This study set out with clear expectation that infants would favour similarity (family resemblance). Although this was weakly supported by the null findings of Experiment 1, it was largely contradicted by the results of the second set of experiments. The mixed pattern of results is difficult to interpret but provides support for the idea that infants’ categorisation can be flexible. The exact categories formed may have depended upon the distribution of features in the category sets or the relative interest and familiarity they held for the infants, but this cannot be established without further research.

**6.1.4. Alien Rabble.** In the final chapter the case was made that greater ecologically validity in tasks would bridge the gap between category learning in the laboratory and the real world. As a step towards meeting this goal, a novel 3D computer game framework was developed. The Alien Rabble game was used in two experiments with 8 year old, 11 year old and adult participants. In one task participants were provided with positive and negative feedback about category members and required to learn either a rule- or similarity-based category structure. In the other task, participants were free to group the exemplars as they wished. In all cases the stimuli were sets of 12 unfamiliar, artificially generated 'aliens'.

In Experiment 1, participants were required to collect positive and avoid negative exemplars in a category learning task. Participants were tested across 6 different
conditions that presented categories based on either a 1-dimensional rule, a 3-
dimensional rule or a similarity structure. Participants of all ages showed learning
in all cases except where categories were based on bimodal similarity structure one
standard deviation apart. Performance on similarity based tasks was comparable
across age groups but adults were markedly more accurate in the rule learning
conditions. Additionally, the adults in the rule learning conditions may have been
responding in a qualitatively different fashion as they showed abrupt improvements
in accuracy suggesting that they may have been learning by means of a rational
hypothesis testing strategy. However, in terms of speed of completing the task
11-year-olds were the fastest group while adults were no faster on average than
8-year-olds.

In Experiment 2, the same groups of participants used a computer interface to
sort a new set of 12 aliens into between 2 and 5 groups of aliens that 'look like
they go together'. As before, the exemplars came from one of six bimodal category
structures. The participants classifications were analysed using a newly developed
measure of dimensional-predictiveness. Analysis showed that in some cases partic-
ipants do detect and classify according to the category structure imposed by the
experimenter and this was more likely in the rule-based categories. Furthermore,
the distribution of uni-dimensional and multi-dimensional sorts was similar across
all three age groups. A control group of 11-year-olds manually sorted printed cards
featuring the same sets of aliens (Experiment 3) and showed a similar pattern of
responses.

One interesting contrast was between the supervised (Exp. 1) and unsupervised
(Exp. 2) tasks. Clear differences were only found between adult and children on
rule-based categories in the supervised experiment, but no differences were found
for similarity based categories or for any of the unsupervised conditions. This
suggests that when required to learn a rational rule-based structure adults can do
so but children have some difficulty. Whereas when given the freedom to classify as
they choose both groups perform in a comparable fashion. Over and above these
findings, the primary conclusion from this work was that the game framework is
a suitable tool for examining similarities and differences in children’s and adults’
category learning abilities. Furthermore, the game has provided a means to improve
the ecological validity of category learning experiments and is useful platform on which to build future research.

6.2. Synthesis

In this section, an attempt is made to draw some broader conclusions from the research conducted in this thesis and to attempt to synthesise and combine some of the findings from the four topics of research. The research was wide ranging taking in participants who ranged in age from 4-months-old to their late fifties and testing them across 9 different experimental paradigms. Given that diversity the conclusions drawn will necessarily be very general. Nevertheless, one advantage of having such a wide ranging approach is that if any generalities are found they could point to deep principles. If any common themes are found across the very different methods and ages that were investigated here, we may hope that any general conclusions would be informative about universal laws. I will limit my conclusions to just two very general themes.

6.2.1. Infants conceptual abilities are grounded in perception. In all the infant studies we have conducted here there is support for the idea of a crucial involvement of low level perceptual abilities in any higher level conceptual abilities that might be identified in early infancy. The most obvious example is from the results of the anticipatory eye movement study in Experiments 1b and 2b in Chapter 3. Several features lead us to believe that the infants' concepts of Same/Different (in this task at least) are perceptually grounded; First, the consistency between the pattern of results found in 4- and 8-month olds, second the continuity with other (pre-conceptual) species and third, the surprising (and not recognisably conceptual) pattern of responses itself. We have already discussed this in Section 3.6.3.

A similar, though weaker, claim could be made about the results of the Cars and Houses study where again there was (1) a common pattern of response between two very different age groups (5- and 11-month-olds), (2) a surprising (and not recognisably conceptual) pattern of responses. Furthermore, even though the pattern of results is 'not recognisably conceptual', that is not sufficient grounds to dismiss the problematic findings of these experiments. Because in both cases there was consistent responding across participants and there was evidence of generalisation.
This could support the interpretation that performance of both groups of infants is being driven by their reaction to the same distribution of perceptual features.

The photographic Same/Different task also appears to tie in with a perceptually grounded explanation. Firstly, if we again compare the results of Wright and Katz (2006) pigeons and monkeys were also able to learn a two item S/D relation with photographic stimuli and in that case performance improved with larger stimulus sets indicating that perceptual experience guided abstraction. Secondly, we may take into account Quinn’s (2003) arguments concerning the perceptual grounding of spatial relation categories, suggesting that only older infants would succeed with richer stimuli.

Finally, in the visual sequence learning task, infants seem to respond to locally 'boring' parts of the sequence by being more likely to look away, here it is not too difficult to quantify 'boring' as a perceptual property of the sequence and it is arguable that the measure of redundancy that we applied to the data does just that. Furthermore, the fact that these are infants’ spontaneous responses suggesting that they driven by infants' immediate reaction to the stimuli, which also suggests a perceptual cause.

Deciding if infants have a conceptual or perceptual basis for learning new categories is important as a general principle because it guides where one looks for explanations and what one accepts as an explanations. It is important to take note of the evidence found here because in other cases the support for perceptual foundations to knowledge may not be so clear cut. Especially, the case when dealing with abstractions like rules where there is often an ambiguity about how dependent the result is on some particular feature of the experiment or if a given concept is sufficient. For example, it might be possible to give an conceptual account of why 8 month old infants succeed at the photographic Same/Different task whereas 4-month-olds fail by claiming that 'the concept of Sameness comes online' somewhere between 4 and 8 months. This explanation may even appeal to other similar concepts which arise at the same time. However, it should be clear that this is really a description rather than an explanation; Whereas, a perceptually grounded account makes a commitment to the idea that concepts are built up from an association of simpler perceptual regularities. Of course, it is still possible that in certain cases certain
concepts do just come online at particular points in development. For example, Spelke (1994) argues that "Initial knowledge may emerge through maturation or be triggered by experience, by [sic] learning processes do not appear to shape it." (p. 439, Spelke, 1994). But all the evidence found here suggests that across a range of different category learning tasks, infants do form conceptual groupings that have their origins in perception.

6.2.2. **Category learning is a complex process.** Taking a perceptual stance when interpreting these results also leads to the conclusion that category learning is a complex process. In this section, I attempt to show that this is more than just a truism.

6.2.2.1. **Category learning is complex.** Despite the fact that we were often looking at very simple category structures with very young and inexperienced participants, it is fair to say that the results were not straightforward. In all of the chapters the results were indicative of complex classifications. In Chapter 2, the infants patterns of response showed some similarities with the group of adult 'controls'. The infants did not passively learn broad global features of the stimuli but appeared to respond explicitly to local redundancy and relative pairwise frequencies. In Chapter 3, both 4- and 8-month-olds showed a surprising pattern of behaviour in the anticipatory eye movement study, learning the dependency for pairs of different coloured shapes, they even generalised this pattern to completely novel coloured shapes. The results of the Cars and Houses experiments in Chapter 4 suggest that infants can be flexible in the types of categories they form depending closely on the exemplars seen; While in the free-sorting tasks in the Alien-Rabble chapter participants of all ages divided twelve items into average of 3.7 groups, indicating that participants were not just performing a binary split of the items. While the results of the *dimensional predictiveness* analysis showed that these groupings were not random but captured the latent structure of the exemplar sets.

All of these results indicate that participants made complex, flexible responses that are highly contingent on the materials seen. Rosch et al. (1976) suggested that the basic level in categorisation serves to simplify the world in a way that reflected its structure. They used a phrase originating in Plato's *Phaedrus* to suggest that categories "carve nature at its joints" (*Phaedrus* 265d-266a). The implication being
that categories are ‘real’ in a Platonic sense, reflecting some true features of the world. Undoubtedly, categories must reflect and simplify the world or they would serve no purpose but the results of the experiments presented here suggest that when infants and adults learn categories they are retaining a substantial part of the complexity and contingency of their individual experience.

6.2.2.2. **Category learning is a process.** Secondly, I believe that it is worth emphasising that category learning is a process. Which is to say that an understanding of the processes involved during category learning is important to understanding the categories formed. Across all the experiments, I believe that there strong support for the claim that one can only understand what happens in the study if you understand the dynamic process by which it unfolds. This is clearly true for the sequence learning results where understanding the time-course of the trials lead to understanding of the infants behaviour. Likewise, the learning process is of interest in the supervised learning task in the Alien Rabble game where adults and children appeared to show different patterns of learning with rule-like category structures. Adults learning quickly and abruptly while children learned more gradually suggesting that adults might be more likely to be learning by means of a rational hypothesis testing mechanism.

It is perhaps harder to see how process is important in free category sorting or in the Same/Different habituation task. It seems like one could explain either of these with a static abstract model, in the first case this might be a static category structure, the second an abstract concept of same/different. However, recall that in the former case different participants gave either unidimensional or multidimensional sorts of the same categories suggesting that there was no single static category structure that describes the behaviour of all participants. In the latter case, we may again refer to the Wright and Katz (2006) result which found that wider experience with more exemplars increased the chances of a Same/Different discrimination in monkeys and pigeons. Furthermore, our other habituation study with photographs of Cars and Houses also suggests a dynamic interpretation. All these results imply that category learning is best understood as a dynamic process. Dynamic models of categorisation are important, this point is expanded upon in the next section.
6.3. Questions for Future Research

It might be appropriate to say nothing further here, as they are already plenty of open questions. A very large number of questions were raised in the introductory chapter and although the following four chapters set out to provide some answers, they did inevitably raise just as many new questions as they solved. As a rule, the specific focused questions directly raised by the individual experiments are better kept in the discussion of those experiments, and the reader is referred back to the general discussion sections of each of those chapters. Then, in the case of the Alien Rabble game, the future work already mapped out constitutes a research proposal in itself and does not need any further elaboration here. I will however focus on one question that I have so far mostly avoided.

6.3.1. The Missing Models. While this doctorate has been a remarkable apprenticeship I feel a slight regret that there was one omission; I did not, in the end, do any modelling. My supervisor, Denis Mareschal, has an international reputation for his use of connectionist modelling as means of understanding early development. He regularly makes a very strong case for the benefits of this neuroconstructivist approach (Mareschal et al., 2007a,b) which I wholeheartedly subscribe to. And as a graduate in mathematics and a former computer programmer, the actual process of building and testing models had tangible appeal. The intention was always to approach this project with a view to using modelling techniques to analyse the results.

As things turned out, I did not get to apply this approach to any the questions I have researched but I continue to believe that the connectionist framework is a very powerful approach to developmental questions. I do not believe in modelling for its own sake and in this case it was more important to conduct the primary empirical research. However, I think that, now that those results have been collected, all of the experiments could well benefit from a neuroconstructivist perspective. Furthermore, it is still possible to take that approach to these results and more to the point I consider this to be the most fruitful avenue for future research that builds directly on these studies. As mentioned above the data from many of my experiments are amenable to modeling directly, this was no coincidence.
In each of the infant experiments we have encountered complex, contingent behaviour that is not easily described by a verbal process theory but could well be amenable to connectionist modelling. It is my belief that something complex is going on each of the infant studies and that explicit modelling may be the best means of understanding the behaviour. In the sequence learning chapter, we have already considered using either Perruchet and Peereman’s (2004) Parser or HAB model of Sirois and Mareschal (2004) to attempt to model the apparent short-term memory effects as the infants see repeating chains of stimuli. If a good model was found this would immediately provide a range of testable predictions for other sequence types.

In the abstract Same/Different learning task with rich stimuli, we speculated that there might be different patterns of looking for (older) infants in the two habituation conditions. Infants in the Same condition appear to switch more as they compare the two identical pictures feature by feature based on their surprising agreement between them, whilst infants in the different condition appear to switch less, dwelling on each whole object for a time. A habituating model with novelty and familiarity effects, such as HAB, could well match this pattern (assuming that there is such a pattern, the full frame by frame coding of the data has not yet been done to confirm this.) The anticipatory eye movement study could also be modelled and the common pattern of responses between ages suggests there is a complex process involved. For reasons of space I will not outline a proposal here but just mention in passing the famous existing model of object occlusion of Mareschal et al. (1999) which may prove an instructive starting point.

Similarly, both the 5- and 11-month-olds in the Cars and Houses experiments in Chapter 4 also showed the same surprising asymmetrical pattern of categorisation. Infants familiarised with cars got bored and stayed bored, while infants familiarized with houses dishabituated to everything. Again, for reasons of space, I will not outline the full reasons why an explicit model may be the best approach, except to observe that there is again a famous prior connectionist model that has accounted for asymmetrical exclusion between cats and dogs (French et al., 2004).

Finally, it is not envisaged that the Alien Rabble data from the current experiments would be sufficient for a learning analysis (Schmittmann et al., 2006) or for fitting
with a neural network model of categorisation (ALCOVE, Kruschke, 1992), but
the framework was intentionally designed to collected large datasets concerning
participants learning of categories with well parametrised, numerically represented
exemplars. And so it is a question of when not if those tasks will be modelled.

6.4. Lessons Learned

In this section, I take several steps back from the research itself to take a wider
perspective on the skills I have acquired, the mistakes I have made and the insights
I gained in the course of past three years. Since many of these revelations will be
of no interest to anyone but myself, I shall try to be brief.

6.4.1. Range wide before drilling deep. One frustration has been that I
have not gone to tremendous depth in any of the topics I have covered. Four very
different strands of research were pursued largely in parallel and all four experimen-
tal chapters were left with unanswered questions and further avenues to be followed
up. In part this is a necessary consequence of the speed with which infancy research
progresses. Any given experiment will take 6 months to turn around so it is wise
to have multiple strings to your bow. The unintended consequence of this has been
that I have not plunged too deep into any one area without an appreciation of the
wider context that all this work fits into. The other sense in which the phrase ap-
plies is in terms of the need to be very well prepared before undertaking an infancy
experiment.

6.4.2. Babies are precious. As any new parent knows, young infants are
precious, rare and grow up too fast. In the scramble for participants, new infancy
researchers discover this too. When you need babies exactly four months old, you
only get one chance to test any given baby. So you should be highly appreciative and
accommodating of the person who has given it to you (so to speak) and completely
prepared to get the most out of your little volunteer.

The greatest moment of terror and dread in my 3 years in the Babylab was when
I discovered the methodological error in the sequence learning Experiment 2. Ten
babies had already taken part and all the time and effort involved in recruiting,
testing and coding them could have been wasted. Fortunately, that data was of
interest and so I escaped at only the price of a great shock and a valuable lesson. Now I take steps to prevent every preventable error and control everything that can be controlled. Though, of course, you will always lose several infants per experiment to fussiness. Infants who will not cooperate but would rather cry. This lead me to another insight I had not fully appreciated before starting to work with infants.

6.4.3. Babies are noisy. Infant data will always have noise because babies will always have more on their minds than just the experiment you are trying to interest them in. Infant research must be all the more exact. So many things can disturb an infant and therefore one needs to control the environment all the more. Extraneous noises and lights can capture a baby’s attention and the wandering attention also forces you to be more creative in keeping their attention and in how you analyse the data. I am coming to believe in rich and engaging experimental set-ups which allow the capture of a rich set of data. A infancy researcher should work with his participants not against them. The same should be said of his colleagues.

6.4.4. Science is collaborative. Prior to my time as a graduate student, I had well understood how science could be competitive and even combative, I had less insight into how it might be collaborative. Three years at Birkbeck’s Centre for Brain and Cognitive Development has changed that. It goes without saying that my research would have been far less productive without the supporting infrastructure of the CBCD. Nor can I quantify the value of the discussion and collaboration with my supervisor. But I do not underestimate the mutual benefits of interaction with wider research group at CBCD and Birkbeck. There is always a cooperative spirit at our annual research network meetings and even in the cut and thrust of conferences, there is a shared sense of curiosity and shared enthusiasm for the questions that draws people together.

6.5. Final Remarks

At the very beginning of this doctorate, the very first book I bought was Greg Murphy’s introductory *The Big Book of Concepts* (Murphy, 2002). At the time, I was charmed and delighted by the title. I could not imagine a more appropriate title for anyone embarking on any Ph.D. The image of being armed with a large almanac of ideas seemed like the perfect starting point where-ever you were going.
But this also emphasised to me that not everyone starting on a doctorate would be handed The Big Book of Concepts. It was a very specialist tome aimed at the narrow field of psychologists and cognitive scientists interested in concept learning. Even someone starting a doctorate in philosophy would not be expected to start out by reading a big book of concepts\(^1\). Yet I also wondered if they ought to be. I would not go as far as Steve Harnad and claim that cognition is categorization (Harnad, 2005). But there is no question that an understanding of the structure of concepts and the process by which they are acquired is part of the foundation of an understanding of cognition. It ought to interest a few of the philosophers too, possibly the epistemologists.

This concept has been a source of joy and inspiration throughout my doctorate. The sense of relevancy of the topic and the sense that I was coming up against a Big Concept, quite literally so, and I was tasked with taking it apart to see how it worked. I felt like I am helping out a very large project that may well be important. Understanding the concept of Concept might be an important problem. It is certainly a big problem. I have just spent three years beginning to understand that.

When faced with a big problem, there is no better place to start than at the beginning. Infancy is the time when we are learning the most and everything that we learn later depends upon the foundations acquired early on. You have to crawl before you can walk. You have to think small before you can think big. So let’s start with the babies. I am very glad I did because, unquestionably, it has been the babies who have been the best part of the whole thing. It has great trying to keep a straight face when explaining infantology to someone new. Telling them in serious tones that

"We do not study babies because we wish to know what babies are thinking about. More often we study babies because we do not know what adults are thinking about. And every adult was a baby once.”

Yes, that’s all true but equally important is that fact that studying babies is about a babillion times more fun than studying adults.

\(^1\)They might perhaps be handed a book of Big Concepts.
My results have proved to me that it may be very difficult to know what infants are thinking about. But that, to a first approximation, unquestionably they are thinking. Their thoughts aren’t easy to understand but that is because they aren’t stupid. They aren’t too smart either but they are learning.

They are always learning.

Small concepts so far but that’s okay because everybody starts small.

"Hello, babies, Welcome to Earth. It’s hot in the summer and cold in the winter. It’s round and wet and crowded. At the outside, babies, you’ve got about a hundred years here. There’s only one rule that I know of, babies—:

'God damn it, you’ve got to be kind."

Kurt Vonnegut Jr. *God Bless You, Mr. Rosewater*, (1965)
Bibliography


Christiansen, M. and Curtin, S. (1999). The power of statistical learning: No need for algebraic


Bibliography


Kant, I. (1781). *Critique of Pure Reason*.


Bibliography


Bibliography 321


