Unsupervised categorisation and cross-classification in humans and rats

James Owen Edward Close

Thesis submitted to

Cardiff University

For the degree of

Doctor of Philosophy

October 2009

DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed .... (candidate)
Date ...30.04.10.........

STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree of ....PhD.............. (insert MCh, MD, MPhil, PhD etc, as appropriate)

Signed .... (candidate)
Date ..30..04..10...........

STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed .... (candidate)
Date ..30..04..10.........

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .... (candidate)
Date....30..04..10........

Acknowledgements

"When darkness is at its darkest, that is the beginning of all light."
(Lao-Tzu, 600 BCE)

This thesis was undertaken in the School of Psychology at Cardiff University. It was funded by a combination of a European Commission Grant 51652 (NEST) and a School of Psychology Studentship.

I would like to thank Lesley-Anne Strabel and Nathalie Walters for their help with the administrative issues that arose during the course of my PhD. I would also like to thank Jeff Lewis, Denis Price and Dennis Simmonds for their assistance with technical issues.

The simple fact is that I would not be where I am now if it were not for my supervisors Ulrike Hahn and Rob Honey. I owe them a great debt of gratitude, not only for their continuous guidance and support, but also for their belief in me throughout my PhD. I would further like to thank Emmanuel Pothos and Dave George, both of whom have been invaluable sources of help and advice at various stages of this thesis. In addition, I would like to thank the 'Hahn Lab' – Adam Corner, Adam Harris, Carl Hodgetts, and Andreas Jarvstad – for their help and useful comments over the course of my PhD.

Finally, the write-up of this thesis has not come easily to me, and there have been many times when 'darkness' has descended. During these periods, however, this 'darkness' has always faded away due to the love and support of my family, and my partner Harriet. Thank you for always being there for me; I cannot wait to get my life back!

This thesis is based on the following publications:

Close, J., Hahn, U., & Honey, R.C. (2009). Contextual modulation of stimulus generalization in rats. *Journal of Experimental Psychology: Animal Behavior Processes, 35,* 509-515.

Pothos, E.M., & Close, J. (2008). One or two dimensions in spontaneous classification: A matter of simplicity. *Cognition, 107,* 581-602.

Thesis summary

This thesis examines how stimulus similarity structure and the statistical properties of the environment influence human and nonhuman animal categorisation. Two aspects of categorisation behaviour are explored: unsupervised (spontaneous) categorisation and stimulus cross-classification. In my General Introduction, I raise the issue of the respective roles of similarity and the classifier in determining categorisation behaviour. In Chapter 1, I review previous laboratory-based unsupervised categorisation research, which shows an overwhelming bias for unsupervised classification based on a single feature. Given the prominent role of overall similarity (family resemblance) in theories of human conceptual structure, I argue that this bias for *unidimensional classification* is likely an artefact. One factor in producing this artefact, I suggest, are the biases that exist within the similarity structure of laboratory stimuli. Consequently, Chapter 2 examines if it is possible to predict unidimensional versus multidimensional classification based solely on abstract similarity structure. Results show that abstract similarity structure commands a strong influence over participants' unsupervised classification behaviour (although not always in the manner predicted), and a bias for multidimensional unsupervised classification is reported. In Chapter 3, I examine unsupervised categorisation more broadly, by investigating how stimulus similarity structure influences spontaneous classification in both humans and rats. In this way, evidence is sought for human-like spontaneous classification behaviour in rats. Results show that humans and rats show qualitatively different patterns of behaviour following incidental stimulus exposure that should encourage spontaneous classification. In Chapter 4, I investigate whether rats exhibit another important aspect of human categorisation; namely, stimulus cross-classification. Results show that the statistical properties of the environment can engender such cognitively flexible behaviour in rats. Overall, the results of this thesis document the important influence of stimulus similarity structure and the statistical properties of the environment on human and nonhuman animal behaviour.

Contents

Tables and Figures

Chapter 0

General Introduction

0. Categorisation in humans and nonhuman animals

Faced with a complex environment, human beings are required to identify efficient strategies to help deal with the world. One important process in this regard is *categorisation*: the assignment of objects, agents, or events to a set of instances of 'the same kind'; for, as noted by Komatsu, "To remember and treat everything in one's environment as unique would require tremendous cognitive capacity" (1992, p. 501). Not only does categorisation provide cognitive economy, it also plays an important role in mediating stimulus generalisation: that is, while classifying two stimuli into the same category will increase generalisation between them, classifying two stimuli into different categories will decrease generalisation between them (Harnad, 1987). Moreover, categorisation allows a person to infer a great deal from only a minimal amount of information (Komatsu, 1992). For example, once a person is informed that a novel entity is a dog, he or she can infer (with confidence) a wide range of different properties about that entity (e.g., that it will bark, chase sticks, etc.). Not surprisingly, then, categorisation forms the foundation for much of human cognition, including higher-level cognitive processes such as reasoning, decision making, and problem-solving.

One notable feature of human categorisation is that it is effortless, irrespective of whether the stimuli are simple, geometric patterns, or complex, naturalistic objects. This effortlessness should be viewed as all the more remarkable considering the incredible flexibility of human categorisation: a single object may be classified at a number of different levels – superordinate (e.g., mammal), basic (e.g., dog), and/ or subordinate (e.g., Labrador; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Tanaka & Taylor, 1991) – and in a variety of different ways depending on the context for classification (e.g., Barsalou, 1982; Tversky & Gati, 1978). For example, while a Labrador may be classified together with a wolf when considering overall appearance, such a classification would seem very odd within the context of "Pets".

Of course, it is not just humans who are faced with a complex environment; nonhuman animals face similar challenges. As a result of these challenges, many different species have been found to engage in complex forms of discrimination learning and 'categorisation' by rote (see Herrnstein, 1990). Indeed, it is likely that these 'basic processes' are integral to everyday functioning and survival, affording efficient generalisation between stimuli. However, while the prowess of discrimination learning in nonhuman animals is undoubted (e.g., Herrnstein, 1979; Herrnstein, Loveland, & Cable, 1976; Vaughan & Green, 1984), fundamental questions have been asked about whether this learning reflects, in any sense, *meaningful*, human-like categorisation (see Chater & Heyes, 1994). Probably the most divisive issue in discussions of the difference between human and nonhuman categorisation is with respect to *concepts*. Following the philosophical distinction between 'intension' and 'extension' of terms first introduced by Frege (1892/1970), 'concepts' are the presumed mental representations that mediate the assignment to 'categories' (classes of objects in the world that somehow 'go together'); or in Murphy's words, "Concepts are the glue that holds our mental world together" (2002, p. 1). In research in humans, the distinction between categories and concepts is often blurred. This blurring is justifiable to some extent because of the critical assumption that categories are the expression of human concepts; when studying human categorisation, we are in effect studying human conceptual structure (at least, that is the assumption). However, this 'blurring' does cause problems during discussions and assessments of categorisation in non-linguistic agents (Chater & Heyes, 1994).

Over the past three decades, a number of influential theories of human conceptual structure have been proposed (a more detailed discussion of these is presented in Chapter 1). First came what has now been termed the "classical view" of concepts, which is premised on the assumptions of *necessity* and *sufficiency* (e.g., Katz, 1972; Katz & Fodor, 1963). That is, categories are assumed to be based on concepts that represent information about the attribute(s) that are necessary *and* sufficient for membership (Komatsu, 1992); consequently, this view has also been termed the "definitional account" of concepts. Born from the philosophical work of Wittgenstein (1953), however, the 1970s brought to the fore a wealth of evidence that ultimately led to the classical view's downfall (see, e.g., Rosch & Mervis, 1975; Rosch et al., 1976; also Fodor, Garrett, Walker, & Parkes, 1980). In its place, new similarity-based views of concepts were soon conceived: first of this kind was *prototype theory*, which assumes that concepts should be considered in terms of a summary (abstracted) representation of category members (Rosch & Mervis, 1975;

see also Hampton, 1995; Posner & Keele, 1968; Reed, 1972). In its strictest form, therefore, category membership of a novel exemplar is determined on the basis of its similarity to, for example, an 'average' dog, an 'average' cat, etc. A second similarity-based approach soon followed in the form of exemplar theory (e.g., Medin & Schaffer, 1978; Smith & Medin, 1981). In exemplar theory, concepts are considered in terms of a set of stored instances, which may or may not be abstracted across during classification (Komatsu, 1992). This view of concepts has been particularly influential, spawning the development of a family of related mathematical models that have provided some of the most detailed modelling of human behaviour to date (e.g., the Context Model, Medin & Schaffer, 1978; the Generalized Context Model (GCM), Nosofsky, 1986; and its connectionist implementation in ALCOVE, Kruschke, 1992; as well as extensions to include reaction times, EBRW, Nosofsky & Palmeri, 1997; and EGCM, Lamberts, 1995, 2000). While fundamentally different in nature, the predictions that arise from prototype and exemplar theory are often indistinguishable. While interesting, this fact has created numerous problems for researchers that are engaged in work comparing the different theories of concepts.

Similarity-based views of our everyday concepts have, however, been attacked on a number of fronts (e.g., Goodman, 1972). These fundamental critiques led to the development of new theories that expounded the theory-like nature of concepts, becoming known as the *theory theory* or *knowledge* approach (e.g., Murphy, 2002; Murphy & Medin, 1985). According to theory theory, concepts are intimately intertwined with people's "naïve theories" about the world: that is, coherent concepts 'fit' with people's general knowledge (Murphy & Medin, 1985). While general knowledge is clearly an important determinant of the way humans behave in their environment, a fully explicated account of theory theory is still to be provided. For example, fundamental questions have not been fully answered: what, exactly, constitutes a theory; how is a theory implemented; how are theories brought to bear in real-world concepts? While some promising suggestions to the answers of these questions have been made (e.g., Kaplan & Murphy, 1999, 2000; Murphy & Allopenna, 1994), it is still the case that the most fully articulated proposals to date of our natural language concepts are prototype accounts (e.g., Hampton, 2001, 2003). Moreover, with respect to attempts to model human categorisation, the computational formalisation of general knowledge effects has proved extremely difficult (see Fodor, 1983; Lewandowsky, Roberts, & Yang, 2006; Murphy, 2002; Pickering & Chater, 1995). Of course, the fact that incorporating effects of general knowledge into models of categorisation is extremely hard does not, in itself, form a platform from which to reject theory-based views of concepts (see, e.g., Heit, 1997, 2001; Heit & Bott, 2000).

By contrast, research in nonhuman animals has typically denied any sense that 'categorisation' in animals is driven by concepts (Chater & Heyes, 1994; but, see Schrier & Brady, 1987, for example). Rather, categorisation-like behaviour is commonly explained in terms of associative principles of learning. This is hardly surprising; the study of concepts in humans is hard enough, given that they can only ever be inferred. Interestingly though, categorisation-like behaviour in nonhuman animals has been explained using theories that are similar in kind to those that have been proposed to explain human categorisation (though in no sense are these theories considered 'theories of concepts'; see Pearce, 1997, for a more detailed overview). For example, similar to the classical view of concepts in humans, feature theory in nonhuman animals assumes categorisation based on a set of defining features (e.g., D'Amato & Van Sant, 1988; Lea, 1984). These defining feature sets, however, are thought to be learned constructs, arising directly from experience. A number of authors have proposed exemplar views of nonhuman categorisation, in which stimulus generalisation provides the mechanism for the successful categorisation of novel stimuli (e.g., Astley & Wasserman, 1992; Pearce, 1988, 1989, 1991). While some authors have tried to claim nonhuman animal categorisation based on a concept (e.g., Schrier & Brady, 1987), simpler learning mechanisms often suffice (e.g., mediated generalisation). As Pearce states, "it remains an open question as to whether or not success by animals in solving any categorisation problems ever implies the possession of a concept" (1997, p. 124; for a similarly critical position, see also Chater & Heyes, 1994). While Pearce's negative conclusion seems justifiable in light of present evidence, it is also apparent that, despite decades of research, the full scope of nonhuman categorisation is still to be resolved. For example, recent work has documented stimulus grouping in nonhuman animals that is beyond the scope of traditional associative analysis (e.g., Honey & Watt, 1998, 1999). One important reason for Pearce's negative conclusion, I would argue, is due to the overwhelming use of *supervised* experimental procedures in investigations of nonhuman categorisation. As has been shown for the study of human categorisation, devoting appropriate resources to the study of *unsupervised categorisation* is essential for the adequate assessment of a species' categorisation ability.

Throughout this thesis, a distinction will be made between supervised categorisation and unsupervised categorisation (see Chapter 1). Briefly, supervised categorisation refers to a situation in which a classifier is required to learn a previously determined classification through trial and error; therefore, feedback is often continually provided. Unsupervised categorisation refers to stimulus classification that proceeds in the absence of feedback, and as such, is assumed to be determined by the classifier's 'natural preferences'. While one may presume that the mechanisms of supervised categorisation drive all classification, this is unlikely to be the case (at least for humans; Pothos & Chater, 2002). First, in humans, there are strong cross-cultural commonalities in the way humans come to categorise the world (e.g., López, Atran, Coley, Medin, & Smith, 1997; Malt, 1995; see Chapter 1). Second, generalisation of new words is often successful following presentation of just a small number of rival category exemplars (e.g., Feldman, 1997). As noted by Pothos and Chater, "This suggests that supervised learning of linguistic categories may be guided by rich prior constraints on what categories are plausible; and unsupervised learning provides a potentially important source of such constraints" (2002, p. 307). To my mind, the importance of research focused on unsupervised categorisation cannot be overstated: simply, it currently allows the best insight into people's 'natural' categorisation biases (or preferences), and our best chance of understanding the fundamental principles that underlie everyday categorisation. Research focused on unsupervised categorisation is particularly important to the debate on whether our natural categories are mainly a product of a structured environment (e.g., Anderson, 1991) or the mind of the human classifier (e.g., Murphy & Medin, 1985)[1]. That is, does the abstract similarity structure for a set of objects bias and guide their classification, or is 'knowledge' (a "naïve theory") about a set of objects most critical in determining categorisation? While certain inferences about natural categorisation can be made from the study of supervised categorisation, the fact that people learn one very specific classification faster than a second very specific classification does not necessarily mean that the first classification is more natural[2].

Until fairly recently, research focused on unsupervised categorisation in humans was rather rare; indeed, even today such work is still dwarfed by research

[1] Of course, it is most likely that categories reflect a complex interplay between the environment and the classifier (see Malt, 1995).

[2] I appreciate that this does not reflect the full scope of supervised categorisation research.

assessing supervised categorisation. Given the inherent difficulty of determining the basis for participant classification in free (unsupervised) categorisation experiments, this is not at all surprising. However, due to the points specified in the preceding paragraph, unsupervised categorisation research has become more abundant. Moreover, a number of influential models of human unsupervised categorisation have now been proposed within the psychological domain (e.g., Anderson, 1991; Love, Medin, & Gureckis, 2004; Pothos & Chater, 2002). While some good progress has been made in our understanding of unsupervised categorisation, one particular anomaly (one might even say perversity) has dominated in findings from laboratory-based unsupervised categorisation research; namely, participants' overwhelming use of a *unidimensional sorting* strategy. That is, when presented with a set of stimuli in the laboratory, people prefer to base their classifications on just one of the *N* stimulus dimensions available (e.g., size). The reason why this is odd is because this does not fit with cognitive scientists' current understanding of natural categories, which conform to the principle of family resemblance (e.g., Rosch & Mervis, 1975; Wittgenstein, 1953). Is there any sense that this bias for unidimensional classification represents a 'natural' preference in human unsupervised categorisation, or is it simply an artefact of the standard experimental setup? Understanding why unidimensional classification is so prevalent within laboratory-based investigations of human unsupervised categorisation is a topic of particular importance, and this issue is the focus of Chapters 1 and 2 of this thesis.

With respect to the dominance of unidimensional classification in the laboratory, some interesting work by Love (2002) has recently highlighted the importance of the distinction between *intentional* and *incidental* unsupervised categorisation (see also, Wattenmaker, 1991). Love (2002) found that whereas intentional unsupervised categorisation was associated with more 'rule-like' (unidimensional) category learning, incidental unsupervised categorisation was associated with more similarity-based categorisation (i.e., a preference for family resemblance structures). This distinction is important because while the majority of laboratory-based unsupervised classification can be considered intentional, natural unsupervised classification will, for the most part, be incidental – that is, stimulus classification will not be the primary objective during a specific interaction (Love, 2002). If one is interested in better understanding the processes that determine natural categorisation, therefore, the focus of experimental, laboratory-based research needs to be on classification that occurs incidentally. Moreover, the development of procedures to assess incidental categorisation in humans will likely prove invaluable for investigations of unsupervised categorisation in non-linguistic beings. Chapter 3 of this thesis picks up on these issues and examines incidental unsupervised categorisation.

To fully understand the roles of the environment and the classifier in determining the formation of categories, one should also look to experiments with nonhuman subjects. If "the mind has the structure it has because the world has the structure it has" (Anderson, 1991, p. 428), then one might imagine that nonhuman unsupervised categorisation would share a number of commonalities with human unsupervised categorisation (Brown & Boysen, 2000): of course, this assumes that nonhuman animals do engage in unsupervised categorisation, and that one has allowed appropriately for issues of scaling, etc. However, as noted above, experimental investigations of nonhuman categorisation behaviour have typically employed supervised classification procedures[3]. This is not surprising; obviously it is not possible to ask an animal to group a set of items in a way that feels natural and intuitive to them, as frequently occurs in studies of human unsupervised categorisation. It therefore remains an open question whether or not nonhuman animals engage in any meaningful form of unsupervised categorisation (of course, this presupposes that animals *do* categorise in some meaningful way under supervised conditions; see Honey & Watt, 1998, 1999). Interestingly, a small amount of work is at least suggestive of the possibility that nonhuman primates have the cognitive requisites to engage in spontaneous categorisation (e.g., Brown & Boysen, 2000; Murai, Tomonaga, Kamegai, Terazawa, & Yamaguchi, 2004; Spinozzi, Natale, Langer, & Brakke, 1999; see also, Spinozzi, 1996). Unfortunately though, this work is less than conclusive: while it documents that some nonhuman primate species do appear to spontaneously recognise the similarity and difference between a set of stimuli, this sensitivity does not, in itself, indicate human-like categorisation behaviour. That is, this work does not show that these animals come to treat a perceptually-based, spontaneous grouping of a set of stimuli in a manner that is truly categorical in nature (cf. Fagot, Wasserman, & Young, 2001).

[3] Interestingly, these supervised classification procedures have documented a number of similarities between human and nonhuman animal categorical discrimination behaviour (see, e.g., Astley & Wasserman, 1992; Wasserman, Kiedinger, & Bhatt, 1988).

Although often not focused on the question of unsupervised categorisation *per se*, a considerable amount of work has assessed how prior, nonreinforced exposure to a set of stimuli influences the way in which nonhuman animals later perceive those stimuli (similar work has also been conducted in human participants; see Goldstone, 1998, for a review). One robust finding from work of this nature in animals has been that nonreinforced preexposure to a set of stimuli results in a later reduction in stimulus similarity (a phenomenon that has been termed *perceptual learning*; see Gibson, 1963, 1969, and, Hall, 1991, for reviews). This result is, of course, the opposite of what one would expect if an animal had come to 'classify together' a set of stimuli (Harnad, 1987). At the same time, following certain other forms of stimulus preexposure in animals, the similarity between two stimuli has been found to increase, which is exactly what one would expect if an animal had come to 'classify together' those stimuli (see Hall, 1991, for a review). This kind of preexposure effect has been termed *sensory-preconditioning* or *acquired equivalence* (Hall, 1991). For the most part, however, sensory-preconditioning has not been considered a product of unsupervised categorisation. Instead, an explanation has been sought on the basis of more simple, associative mechanisms (e.g., Hall, 1991; but see, Bateson & Chantrey, 1972; Chantrey, 1974).

Whether or not nonhuman animals engage in unsupervised categorisation is a fundamental question to assessments of animal cognition, and yet there is a paucity of research which has adequately addressed this. Not only is it important to establish if unsupervised categorisation is an *evolutionary primitive*, but if it is, then research of this kind gets to the heart of the debate on the role of the environment versus the classifier in unsupervised classification. Consequently, as well as focusing on incidental unsupervised categorisation in humans, Chapter 3 of this thesis also investigates the possibility of incidental unsupervised categorisation in rats.

That unsupervised categorisation (or learning) may take place with respect to the statistical properties of the environment finds a natural home within associative and connectionist analyses of learning. Interestingly, while some connectionist models limit the flexibility of nonhuman animal behaviour and deny the formation of 'internal representations' (e.g., Pearce, 1994), this is not true of other, more complex connectionist architectures. Indeed, those connectionist networks that typically employ *hidden units*, such as Elman's (1990) simple recurrent network (SRN), are, in a sense, capable of "building complex internal descriptions" (Gureckis & Love, in press, p. 6; see Elman, 1991; Rumelhart, Hinton, & Williams, 1986; also, Mareschal & Quinn, 2001). Such architectures have now been used widely to model both supervised (e.g., Gluck & Bower, 1988; Kruschke, 1992) and unsupervised categorisation in humans (e.g., Japkowicz, Myers, & Gluck, 1995; Love et al., 2004; see also, Japkowicz, 2001). Moreover, they have been employed to account for categorisation behaviour in nonhuman animals that is beyond the scope of traditional associative theory (e.g., Honey & Ward-Robinson, 2002; see also, Honey & Watt, 1998, 1999). What these more complex connectionist architectures demonstrate, therefore, is that simple associative mechanisms can afford a surprising degree of cognitive flexibility (this is further reinforced by "hybrid" models of associative learning; e.g., Le Pelley, 2004). Naturally, this has a number of important implications with respect to nonhuman animal categorisation: that is, it should be considerably more flexible than once assumed by traditional learning theory (e.g., Pearce, 1994; Rescorla & Wagner, 1972). If correct, then this would be of notable interest to discussions of the differences that might exist between nonhuman and human categorisation, and nonhuman categorisation *per se* (e.g., Chater & Heyes, 1994). Taking up this theme, Chapter 4 explores one prediction made about the cognitive flexibility of nonhuman animals, based on a connectionist analysis outlined by Honey and Ward-Robinson (2002).

The fact that connectionist networks, based on the principles of associative theory, have shown promise in modelling both supervised and unsupervised categorisation is very exciting. Indeed, connectionist architectures are perhaps best placed to offer not only a single model of categorisation that unifies supervised and unsupervised classification in adult humans (see Love et al., 2004), but also a single model of categorisation that tracks the course of human development and unifies human and nonhuman classification. While still a long way off, the success of connectionist architectures to date appears to show some genuine promise in this regard (see, e.g., Mareschal & Quinn, 2001). However, to fulfil that promise, closer collaboration between researchers that study human and nonhuman categorisation will be necessary. One of the overarching aims of this thesis, therefore, is to draw liberally from both of these rich domains, assessing aspects of human *and* nonhuman animal categorisation behaviour.

Chapter 1

Unsupervised (spontaneous) categorisation: A Review

1. The nature of human unsupervised categorisation

Why do we have the categories we have? That is, given the almost infinite number of different ways that humans could divide up the world, why have we come to form the category of objects that we now refer to as dogs, and not a category of objects that includes dogs, oak trees and computer speakers? As highlighted in Chapter 0, *categorisation* – the grouping together of a set of entities that are regarded to be 'alike' in some way – is both powerful and flexible; it affords an agent the ability to efficiently identify, reason and infer properties about objects in the world (including objects not seen before).

To answer the question of why we have the categories we have and not others, it is important to understand the mechanisms that guide categorisation. In a broad sense, should categorisation predominantly be regarded as a product of the *human mind*, in which resides language and general knowledge about the world (Murphy & Medin, 1985), or predominantly as a product of *structure in the environment* (in the form of perceived regularities and discontinuities; Rosch & Mervis, 1975)? "Predominantly" here reflects the fact that it is unlikely that these factors are mutually exclusive; rather, the role of the categoriser will likely interact with environmental structure (Malt, 1995). For example, despite two mushrooms being perceptually similar, 'knowledge' may impact upon one's decision to eat both mushrooms for dinner: whereas one of these mushrooms is an edible straw mushroom (*Amanita virgata*), the other is actually a highly poisonous Death Cap mushroom (*Amanita phalloides*). Of course, unless one is a mycologist, it is unlikely that one would have the knowledge necessary to tell these two mushrooms apart. In human language, there is "*division of linguistic labor*" (Putnam, 1996, p. 287). That is, it is not necessary for everyone to be able to make the distinction between the straw mushroom and the Death Cap mushroom, despite this distinction being important to everyone. However, the words 'straw mushroom' and 'Death Cap mushroom' would be meaningless unless someone was able to distinguish between these two types of mushroom (Putnam, 1996).

The upshot of this is that words alone do not create meaning (Putnam, 1996): For example, if I were to start calling all dogs 'blibs', I would still 'mean' *dog* when I said *blib*: the entities that I now refer to as *blibs* have not changed in any respect from when I called them *dogs*. At the same time, my communication with others about these objects will, of course, be detrimentally affected, until that is they realise that I 'mean' *dog* when I say *blib*. Critically, once this has been established, effective communication can then resume. One way in which effective communication may be re-established is if I, and the person that I am speaking to, share a similar *concept* of 'dogness'; that is, what it means to be a dog (or 'blib'; e.g., they have fur, bark, etc.). As noted in Chapter 0, a concept is commonly regarded to be the minimal unit of information (in the head) that is required to determine a categorisation (i.e., the physical (external) grouping of stimuli). While early theories of concepts did assume a primary role for language – such that knowing the definition of a word meant that you had mastered the concept of the object being defined (see Komatsu, 1992) – this view was challenged on a number of fronts by the philosophical work of Wittgenstein (1953), and the empirical work of Eleanor Rosch and her colleagues (e.g., Rosch & Mervis, 1975; Rosch et al., 1976) on basic level categorisation. As will be detailed over the course of this review, Rosch's work spawned the development of new theories of human concepts based on *similarity*. This change in focus was not only important with respect to human concepts and categorisation, but it also had implications for comparative assessments of human and nonhuman animal categorisation. This is because categorisation based on overall similarity, as opposed to rule-based cognitive processing (i.e., through definitions), is consistent with the use of associationistic processes, which all animals are thought to share (Lea & Wills, 2008). However, as will also be shown, similarity-based views of human conceptual structure have faced fundamental critiques of their own (e.g., Murphy & Medin, 1985).

1.1 Theories of human conceptual structure

As highlighted in Chapter 0, a number of influential theories of human conceptual structure and *category coherence* (that is, what makes categories 'good') have been proposed over the past three decades (see Smith & Medin, 1981). Using the terminology of Komatsu (1992), these theories can be broadly broken down into similarity-based views (e.g., Hampton, 1979; Katz, 1972; Lakoff, 1987a, 1987b;

Medin & Schaffer, 1978; Nosofsky, 1984, 1986, 1988a, 1988b, 1991; Reed, 1972; Wittgenstein, 1953) and knowledge-based views (Murphy & Medin, 1985, see also, Medin & Wattenmaker, 1987).

1.1.1 The Classical View

The *classical view* (formalised by Katz, 1972; see also, Katz & Fodor, 1963) assumes that concepts specify individually necessary and jointly sufficient constraints for categorisation: that is, concepts are regarded as definitional in nature. Critically, the classical view of concepts is intimately intertwined with natural language, such that having definitional information about a word assumes possession of the defined concept (Ogden & Richards, 1956; see also, Gleitman, Armstrong, & Gleitman, 1983; Komatsu, 1992; Medin & Smith, 1984). Importantly, this definitional information is considered distinct from encyclopaedic information (i.e., information about how category members relate to other aspects of the world; Komatsu, 1992). The assumption of necessity and joint sufficiency of attributes means that the classical view is extremely rigid, implying an account of conceptual structure in which category membership is clear-cut and discrete. That is, an object X either *is* or *is not* a member of category Y, and no category member can be more or less typical of a category than any other category member (Komatsu, 1992). Consequently, categories will always be maximally coherent, and should follow rule-like structures such as "if a stimulus X has a square head then classify as a member of Category A, else classify as a member of Category B". The classical view does not, however, address why some categories are preferred over others (i.e., the "naturalness" of certain categorisations over others). Indeed, any arbitrary categorisation that fits within the definitional constraints of the classical view would be considered as 'natural' as any meaningful categorisation (Komatsu, 1992).

As noted in Chapter 0, the classical view has an obvious analogue in the animal learning literature; namely, feature theory (e.g., D'Amato & Van Sant, 1988; Lea, 1984). Here, categorisation is assumed to be guided by a set of learned features that define whether a stimulus is positive (i.e., has been associated with reward) or negative (i.e., has been associated with the absence of reward). Based on this theory, therefore, nonhuman animal categorisation is assumed to be determined by necessary and jointly sufficient features.

1.1.1.1 The rejection of necessity and joint sufficiency in concepts

Born from the philosophical work of Wittgenstein (1953), which advocated the principle of family resemblance in categorisation, researchers in the 1970s began to attack the assumptions made by the classical view (see Mervis & Rosch, 1981; Rosch, 1978; Smith & Medin, 1981). Leading this attack was work undertaken by Rosch and her colleagues (e.g., Rosch & Mervis, 1975; Rosch et al., 1976) on basic level categorisation. In a series of experiments, Rosch et al. (1976; see Mervis & Rosch, 1981, for a review) observed that, when categorising a set of items, humans normally consider one level of abstraction to be more 'natural' than others. For example, when making category judgements about a set of dog stimuli, participants will normally be faster and more accurate to respond that a particular stimulus is a dog (the *basic level* of abstraction) compared to a Labrador (the *subordinate level* of abstraction) or a mammal (the *superordinate level* of abstraction). Some authors have argued that the basic level of categorisation reflects inherent structure within the environment, and experiments using structured, artificial taxonomies lend some support to this view (i.e., by ruling out other factors such as background knowledge, etc.; e.g., Lassaline, Wisniewski, & Medin, 1992; Murphy & Smith, 1982).

Research on basic level categorisation has obvious conceptual links with unsupervised categorisation in that it has sought to identify what it is that makes categories at the basic level 'good' categories (see, Corter & Gluck, 1992; Gosselin & Schyns, 2001; Jones, 1983). The findings of Rosch and her colleagues (see Mervis & Rosch, 1981), therefore, are important in the context of unsupervised categorisation in a number of respects: First, they have shown that 'real world' category structures are broad, rich constructions that are patently not based on definitional features. For example, when asked to list the features of members of a particular category, people produce a broad spectrum of answers, rather than all focusing on a core set of necessary features (e.g., Rosch & Mervis, 1975; Rosch et al., 1976). When studying category construction in the laboratory, therefore, one would presume that the formation of categories will similarly adhere to these 'natural' principles (i.e., that they will be based on family resemblances). Second, when based on perceptual similarity, at least, people's category construction should reflect the perceived similarity-based regularities that exist between a given set of items, such that categories maximise within-category similarity and minimise between-category similarity (Pothos & Chater, 2002).

People do not *always* prefer the basic level, however: for example, studies have shown faster and more accurate verification at the subordinate level of abstraction for atypical category members (e.g., a penguin, Murphy & Brownell, 1985) and among experts of a specific category (e.g., birdwatchers and dog experts, Tanaka & Taylor, 1991; see also, Berlin, Breedlove, & Raven, 1973; Dougherty, 1978). Similarly, Murphy and Wisniewski (1989) found matched verification times for both basic and superordinate level categorisation when the items to be categorised were presented within a specific context (e.g., judging a picture of a chair to be a 'chair' or 'furniture' when presented in a living room scene). Recently, faster and more accurate verification at the superordinate level of abstraction has also been shown among participants engaged in a speeded categorisation task (Rogers & Patterson, 2007). The fact that humans view one level of abstraction as 'most natural', regardless of what level that is, does not fit with the assumptions of the classical view. Further evidence amassed against the classical view has included results showing that category boundaries are not clear-cut and discrete, but rather fuzzy (e.g., McCloskey & Glucksberg, 1978; Hampton, 1979, 1981), with some category members being more or less typical of a category than others (e.g., Lakoff, 1972; Rosch, 1973, 1975; Rosch & Mervis, 1975). Moreover, from an intuitive viewpoint, category formation based on necessary and jointly sufficient features appears unable to capture the breadth and depth of human category structures (e.g., Rosch & Mervis, 1975). Ultimately, these criticisms led to the rejection of the classical view (see Fodor, Garrett, Walker, & Parkes, 1980).

1.1.2 The rise of alternative theories of conceptual structure: similarity-based or theory-based?

With the rejection of the classical view came the rise of probabilistic accounts of conceptual structure, invoking a central role for similarity. Most prominent within these probabilistic theories are the *family resemblance* (or *prototype*) *view* (see, e.g., Hampton, 1979; Reed, 1972; Wittgenstein, 1953) and the *exemplar view* (see, e.g., Medin & Schaffer, 1978; Nosofsky, 1984, 1986, 1988a, 1988b, 1991). These similarity-based views contrast with an alternative view of conceptual structure proposed by Murphy and Medin (1985, see also, Medin & Wattenmaker, 1987), termed *theory theory.*

1.1.3 The family resemblance (prototype) view of conceptual structure

The family resemblance view focuses on the relationship of the elemental overlap between a set of items. That is, within a category, items will share a high level of commonality (in terms of number of shared elements); between categories, items will share a lower level of commonality. Furthermore, those category members that share a great deal of overlap (or family resemblance) with other members of the category will be viewed as more prototypical of the category as a whole (Rosch & Mervis, 1975). Taken in its strictest form, one could propose that each category is represented by a single, ideal member, sometimes termed the category prototype. However, this position is not one that has been widely accepted within the literature (Murphy, 2002). Such a view would not allow for information about category variability, and as Murphy (2002) notes, it is difficult to conceive of a single prototype that could represent an "ideal bird", for example. Instead, theorists have conceptualised the prototype view as the formation of a summary representation of a category as a whole (Hampton, 1979; Smith & Medin, 1981). With the introduction of feature weightings, these summary representations allow for considerable complexity, enabling representations to convey such information as the variability of a category and the typicality of a specific category member (Murphy, 2002). For example, the feature 'barking' may be weighted more highly than the feature 'having four legs' in one's representation of a dog, since many animals have four legs, and so captures little of the uniqueness of what it means to be a dog.

Prototype accounts of categorisation have not found a natural home within the animal learning literature. This is not particularly surprising; abstraction across a category as a whole – to allow for the formation of a summary representation – implies, in humans at least, the formation of a concept. As noted in Chapter 0, however, animal learning theorists have typically denied any sense of concept formation in nonhuman animals (see Chater & Heyes, 1994). In support of this, the classic 'prototype effect' found in humans has often not been found in nonhuman animals (e.g., Lea & Harrison, 1978). The 'prototype effect' describes the observation that stimuli that share a high degree of similarity with a category prototype are classified more readily than stimuli that are quite different from the category prototype (e.g., Posner & Keele, 1968). While some authors have now demonstrated such behaviour in nonhuman animals (e.g., in pigeons; Aydin & Pearce, 1994), as Pearce (1997) points out, feature (McClelland & Rumelhart, 1985) and exemplar theories (Shin & Nosofsky, 1992) appear to offer an account of the 'prototype effect' without needing to appeal to the formation of prototypical (summary) representations.

1.1.3.1 The family resemblance view and its critics

Born from the failures of the classical view, it is of little surprise that the family resemblance view is readily able to explain such phenomena as typicality effects; indeed, this view predicts these effects. However, a number of shortcomings of the family resemblance view have been highlighted over the last ten years. It is intriguing, for example, that humans often hold deep beliefs that necessary and sufficient conditions do form the basis for categories, even if they cannot express what these are (e.g., McNamara & Sternberg, 1983). Furthermore, some authors have questioned whether typicality effects in themselves provide adequate grounds for the rejection of the classical view and the acceptance of the family resemblance view (Armstrong, Gleitman, & Gleitman, 1983; Gleitman et al., 1983). Other problems for the family resemblance view include a loss in power for explicating linguistic meaning, inductive reasoning, and the formation of complex concepts (e.g., 'salmon fillet' from the concepts 'salmon' and 'fillet'; Komatsu, 1992). Lastly, with the naturalness and coherence of concepts relying on an interaction between the environment and the classifier's perceptual system, it has been argued that this view can only account for perceptually based concepts; if true, this would clearly make it an inadequate view of adult concepts (Neisser, 1987).

1.1.4 The exemplar view of conceptual structure

Compared to the family resemblance view, where a concept is regarded as a stored abstraction of a category as a whole, the most widely suggested form of the exemplar view – the *instance* approach – takes, essentially, the opposite viewpoint. That is, a concept reflects sets of different, individually stored exemplar representations, with abstraction across a category only occurring during concept use (Komatsu, 1992). However, this is not to say that abstraction always, or even ever, occurs when using a concept. If no abstraction takes place, then this suggests the intriguing possibility that, "In some sense, there is no real concept (as normally conceived of), because there is no summary representation" (Murphy, 2002, p. 49). The exemplar approach has proved to be an extremely successful account of human categorisation, spawning some of the most detailed and successful modelling to date. One of the most influential models of this form is the Generalized Context Model (e.g., Nosofsky, 1986, 1988a, 1989; Nosofsky, Clark, & Shin, 1989; see also, Kruschke, 1992). Although this model has generally been applied to assessments of categorisation where feedback is provided, recently, it has also been applied to assessments of categorisation where feedback is not provided (Pothos & Bailey, 2009). Along with feature theory, a number of authors have applied the instance approach of exemplar theory to nonhuman animal categorisation to good success (e.g., Astley & Wasserman, 1992; Pearce, 1988, 1989, 1991). As noted in Chapter 0, stimulus generalisation provides the additional mechanism to explain humans' and nonhuman animals' ability to accurately classify novel stimuli.

1.1.4.1 The exemplar view and its critics

Komatsu (1992) has argued that the exemplar view does not provide a strong account of why some groupings are privileged over others (i.e., category coherence). He states that "With no prior specification of the nature or degree of similarity necessary for items to be instances of the same concept, there is no constraint at all on possible new instances: At the very extreme, every object is similar to every other object in some way" (Komatsu, 1992, p. 509). Here we see Komatsu echoing the thoughts of Goodman (1972), who argued that to say two things are similar (and so be classified together) without qualifying in what respects the two objects are similar, is a vacuous statement devoid of content. For example, my computer and my desk are similar in an infinite number of ways; they both weigh less than one ton, two tons, three tons etc. I therefore need to qualify the similarity relationship between the two objects by saying how they are similar: my computer and my desk are similar in *respect* to the fact that both items can be found in my office. However, once one has introduced the notion of 'respects' into similarity judgments, Goodman (1972) argues, then it is these 'respects' that take on all the explanatory power, leaving no role for similarity *per se*. As such, one might argue that similarity is too unconstrained to afford the constrained nature of categorisation. Without extra specification, therefore, it seems the exemplar view of conceptual structure is deeply flawed.

More recently, however, the arguments of Goodman (1972) have been challenged on two fronts: First, when presented with objects formed from multiple dimensions – as would always be the case in the real world – 'respects' can only do some of the work, as the central issue psychologically is how different dimensions are combined to form an overall similarity judgment (Goldstone, 1994; Hahn & Chater, 1997; Medin, Goldstone, & Gentner, 1993). Second, when dealing with the notion of similarity within object categorisation, some authors have argued that the focus should centre on an agent's mental representation of an object, and not on the objective properties of the object *per se*. By their very nature, it is argued, mental representations will be representative only of those dimensions that are important to successful classification, and so by extension, finite (Hahn & Chater, 1997). Consequently, those dimensions which are clearly arbitrary to effective classification (such as weighs less than 1 ton) will not be represented; and so, similarity is naturally constrained in object categorisation. If a person is presented with two objects that are highly dissimilar, therefore, then a person's finite representations of these objects will likely contain no similarities whatsoever; as such, these objects would not be classified together.

Over the course of experience, it is also probable that the similarity relations between objects will naturally start to shift into some coherent pattern. That is, a category will cohere by virtue of the fact that instances within a category will share a greater degree of similarity to each other than instances from different categories. Consequently, as for the family resemblance view, one is left with a situation in which coherent categories will reflect some optimal ratio between maximising within-group similarity and minimising between-group similarity (Rosch, 1975). Certain classifications, such as those at the basic-level, will therefore be privileged because they best attain this 'optimal' ratio (see Rosch et al., 1976).

More problematic for an exemplar view of conceptual structure that invokes dimensional summation as its metric of similarity (see also, Tversky, 1977) is that such an account appears unable to reflect more sophisticated, relationally-based forms of similarity, and the subsequent categorisation this affords. Indeed, it is more than possible that a dimensional summation strategy is only intuitive for binary-valued stimulus structures. However, alternative metrics of similarity can help to overcome this problem (see, e.g., Hahn, Chater, & Richardson, 2003; Markman & Gentner, 1993). In the unsupervised categorization experiments presented in Chapters 2 and 3 of this thesis, all category structures are specified along continuous valued dimensions (e.g., size).

1.1.5 Theory theory

Dissatisfied with the notion that similarity-based views of conceptual structure provide a principled account of conceptual coherence, Murphy and Medin (1985; see also, Keil, 1989) proposed an account of concepts based on theoretical knowledge. Instead of regarding similarity as the sole basis for conceptual coherence, they propose that it is people's *theories* about the world that provide the 'glue' that holds concepts together. What, then, is a theory in respect to concepts? First, Murphy and Medin (1985) do not use theory to mean a scientific account of something. Instead, they use it to mean any of numerous mental "explanations" that generally take the form of causally connected sets of relations between concepts (Murphy & Medin, 1985). That is, concepts cohere to the extent that knowledge is available which causally relates the instances of a category: The stronger the relations are that link together the instances within a category, the greater the category coherence of that category. For example, the concept "BIRD" can be considered very coherent because the theoretical knowledge 'most birds fly' and 'wings afford flight' supports the high feature correlation between birds and 'having wings'. But how can theories explain concepts when concepts are made out of theories; clearly, this is circular? Murphy and Medin embrace this circularity, arguing for a bidirectional influence between concepts and knowledge: "Concepts and theories must live in harmony in the same mental space; they therefore constrain each other both in content and in representational format" (1985, p. 313).

It is important to note that Murphy and Medin regard the knowledge approach as supplying the constraints that are missing from similarity-based views of conceptual coherence, rather than seeing it as a purely contradictory account of such. Numerous variants of the knowledge approach to conceptual coherence now exist: these include, for example, *psychological essentialism* (e.g., Medin & Ortony, 1989), *idealised cognitive models* (Lakoff, 1987a, 1987b), and *mental models* (Johnson-Laird, 1983; see Komatsu, 1992). Due to its focus on 'naïve theories' about the world constraining categorisation, the theory theory (or the knowledge approach; Murphy, 2002) places a strong emphasis on the role of the classifier in determining this process. As a consequence of this, it is apparent that human categorisation will most likely be *qualitatively*, rather than simply *quantitatively*, different from nonhuman categorisation. Whether or not nonhuman animals have any kinds of 'naïve theories' about the world is a fascinating question in its own right; whatever the answer, however, it seems clear that these theories will not be comparable to those of humans.

1.1.5.1 Theory theory and its critics

While 'naïve theories' ('knowledge') about the world may play an important role in human categorisation, by guiding which categories appear coherent, the theory itself remains rather underspecified. As noted in Chapter 0, fundamental questions remain with respect to how theories are implemented and brought to bear in real-world concepts (Close, Hahn, Hodgetts, & Pothos, 2009). Moreover, the difficulties posed in incorporating general knowledge factors into models of human categorisation are well known; indeed, these difficulties may prove insurmountable (cf. Fodor, 1983; Lewandowsky et al., 2006; Murphy, 2002; Pickering & Chater, 1995; but, see Heit, 1997, 2001; Heit & Bott, 2000). Given this state of affairs, it is still the case that the most fully articulated proposals of natural concepts are those based on similarity (e.g., Hampton, 2001, 2003; Nosofsky, 1986). It is no wonder, therefore, that probabilistic accounts of concepts (e.g., prototype and exemplar views) still engender wide support.

In summary, within a culture, and even between cultures, the category structures formed by humans are often similar (Malt, 1995). While some authors have argued that this observation reflects the impact of environmental factors constraining category construction (e.g., Billman, 1989; Malt & Smith, 1984; Rosch & Mervis, 1975; see also, Anderson, 1991; Smith & Heise, 1992), others have promoted a view of category coherence driven simultaneously by low- and higher-level cognitive processes (e.g., Murphy & Medin, 1985; Wattenmaker, Dewey, Murphy, & Medin, 1986). These viewpoints have been expressed over the past three decades in a number of influential theories of human conceptual structure, which have been outlined above. Despite notable advances in our understanding of human categorisation, however, there still exists no unified position on which theory of conceptual structure, if any, is correct. The fact is, categorisation is most likely a product of both the environment and the classifier (see Malt, 1995): but the question remains, which factor dominates? In the subsequent review, I look to assess what research focused on laboratory-based *unsupervised categorisation* tells us about the nature of human conceptual structure. Specifically, what insights does unsupervised categorisation research provide about the factors that guide and determine everyday category construction? Moreover, wherever pertinent, I will seek to draw parallels between findings from unsupervised categorisation research in humans and findings from research in nonhuman animals. Such comparative assessment is both interesting and important, as it allows for a more accurate assessment of the role of the classifier in determining categorisation behaviour.

1.2 Supervised versus unsupervised categorisation

To reiterate from Chapter 0, the study of human categorisation (and to a far lesser extent nonhuman categorisation), has been pursued in two distinct contexts. The first is where participants are asked to discover or impose a classification on a set of unlabelled objects. This is undertaken without any feedback on performance, and as such, has been termed *unsupervised* categorisation. The second is where feedback is typically continuously provided to participants; consequently, it has been termed *supervised* categorisation. Here, a participant's task is to learn a predefined classification from a set of labelled instances as quickly as possible. This conventional distinction between unsupervised and supervised categorisation will be adhered to throughout this thesis.

1.3 Unsupervised categorisation: informing understanding of human conceptual structure

While the majority of categorisation research to date has investigated supervised categorisation, more recently, the merits of unsupervised categorisation as a tool for assessing human conceptual structure have been realised. As noted in Chapter 0, the logic runs as follows: it seems reasonable to suppose that the categories people prefer to construct when provided with no supervision will reflect those mechanisms that underlie people's real world (natural) categorisations. Therefore, in comparison to tasks that are supervised in nature, unsupervised categorisation allows for an assessment of the principles governing human conceptual structure in an unconstrained manner; hence the reason it is the focus of this review. This is not to deny the utility of research that has employed supervised procedures. Indeed, some of the most detailed modelling of human behaviour to date has benefited directly from such work (e.g., Kruschke, 1992; Lamberts, 1995, 2000; Medin & Schaffer, 1981; Nosofsky, 1986; Nosofsky & Palmeri, 1997).

1.3.1 Necessary and jointly sufficient features in unsupervised categorisation

In a seminal paper by Medin, Wattenmaker and Hampson (1987), participants were found to display an overwhelming preference for *unidimensional classification* when asked to sort a set of stimuli constructed from an apparently intuitive family resemblance structure. The stimuli used by Medin et al. (1987) included pictures and phrases, where each dimension represented either a certain attribute or phrase, respectively. Category A was formed from the prototype 1, 1, 1, 1, and Category B was formed from the prototype 0, 0, 0, 0, with the other items of a category differing from their respective category prototype by a single feature (see Figure 1). Similarly, using stimuli that consisted of two dots depicted on pieces of white card, which varied in interdot distance, orientation and overall position, Imai and Garner (1965) showed that participants predominantly chose to base their classifications on only one of the three dimensions available, rather than on all three dimensions. These findings would suggest, therefore, that participants are predisposed to employ a categorisation strategy based on unidimensional rules; that is, based on the principle of necessity and joint sufficiency. For example, in Medin et al.'s (1987) Experiment 1, classification of cartoonlike animals into Category A might be defined by the *necessary* presence of four legs, whereas classification into Category B would be defined by the *necessary* presence of eight legs.

Family resemblance sort				
Category	D1	D2	D3	D4
a	1	1	1	1
a	1	1	1	0
a	1	1	0	1
a	1	0	1	1
a	0	1	1	1
b	0	0	0	0
b	0	0	0	1
b	0	0	1	0
b	0	1	0	0
b	1	0	0	0

Figure 1. The abstract stimulus structure employed by Medin et al. (1987) in their Experiments 1, 2b, 2c, 2d, 3, and 6. The first column indicates the assumed optimal classification of the items, when considering all four dimensions of variation together (this reflects the family resemblance classification). D1 – D4 represent individual stimulus dimensions; for example, head shape, number of legs, body markings, and tail length. These dimensions can take a value of 0 or 1, where a value of 0 on D1 reflects an angular head and a value of 1 on D1 reflects a round head, for example. In boldface are shown the assumed prototypes of each category (in four dimensions).

More recently, Regehr and Brooks (1995) assessed the impact of a number of task manipulations on participants' preference for unidimensional classification. These manipulations included the following: increasing the family resemblance structure of a stimulus set through the addition of more dimensions (see also Medin et al., 1987); making the stimuli appear more integral by decreasing the separability of their dimensions of variation; and, providing a simple rule that defined the two category, family resemblance structure. In line with findings by Medin et al. (1987), Regehr and Brooks (1995) found that none of these manipulations had any impact on reducing participants' preference for unidimensional classification. Building on this work, Milton and Wills (2004) have further shown a bias for unidimensional classification among participants engaged in a sequential 'matching-to-standards' procedure (in contrast to other findings by Regehr & Brooks, 1995). The sequential matching-to-standards procedure takes the following form: following an initial pre-sort phase – where participants are simply asked to group together identical pairs of stimuli – two prototypes are placed side by side on a table. Participants are told that these two prototypes are characteristic of Category A and Category B, and they remained visible on the table throughout the experiment. Half of the stimuli given in the pre-sort phase are then presented to participants, and they are asked to place each stimulus into the category of their choosing (Category A or Category B). Participants group the stimuli in a sequential fashion, and they are told to place each stimulus card face down directly below the group they feel it most resembles (see Milton & Wills, 2004; Milton, Longmore, & Wills, 2008). By employing this sequential matching-to-standards procedure, Milton et al. (2008) have also documented evidence that under conditions of high time-pressure and concurrent cognitive load, participants' preference for unidimensional categorisation is increased relative to conditions of low time-pressure and no cognitive load. This finding is surprising given previous results by Ward (1983, discussed later; see also Smith & Kemler Nelson, 1984; Ward, Foley, & Cole, 1986), but is in line with accounts of categorisation based on stochastic sampling (e.g., Lamberts, 2002) and dimensional summation (Milton & Wills, 2004).

More evidence for a unidimensional classification preference among participants comes from work by Ashby, Queller, and Berretty (1999; see also, Fried & Holyoak, 1984; Homa & Cultice, 1984). They found that, in the absence of any feedback, participants were unable to learn an experimenter defined category structure when the boundary separating the contrasting categories was orthogonal. Indeed, when analysing their participants' behaviour, Ashby et al. (1999) concluded that their failure to learn the orthogonal category structure resulted from them trying to impose unidimensional or conjunctive rules, rather than employing a classification strategy based on an integration of both dimensions of variation. When feedback was provided, participants were able learn the category structure specified along the diagonal. In contrast, when the categories were separated by a unidimensional boundary, the experimenter defined category structure was readily learned both when feedback was available, and when it was not (a depiction of the category structures employed by Ashby et al., 1999, is presented in Figure 3; see Chapter 2). Ashby et al.'s (1999) stimuli were lines that varied in both length and orientation, and learning was assessed by monitoring participants' increased levels of categorisation accuracy over 800 trials.

The above findings show that unidimensional classification in laboratory-based unsupervised categorisation tasks is both highly pervasive and robust. Indeed, Ashby et al. even go so far as to say that "in the absence of feedback, people are constrained to use unidimensional rules" (1999, p. 1). Ahn and Medin (1992) have similarly concluded that unidimensional classification is a ubiquitous feature of unsupervised category formation, invoking such in their two-stage model of category construction. From the standpoint of many categorisation researchers, however, the results presented above are quite odd. Simply, they do not fit with current understanding of the nature of everyday (natural) categories in humans, which are patently not definitional in kind. That is, with respect to everyday categorisation at least, unidimensional classification must be considered suboptimal to classification based on a principle of family resemblance. The reason for this is that, in a complex environment, classification based on such a restricted, definitional principle would simply be too inflexible. In contrast, classification based on family resemblances would afford great flexibility. For example, if one took the presence of a wing as a definitional attribute for "birdness", then one would have to wrongly classify a bat as a bird. Similarly, if one took the presence of a blow hole as a definitional attribute for "whaleness", then one would have to wrongly classify a dolphin as a whale. Moreover, as noted earlier, the idea that everyday concepts are based on necessary and jointly sufficient features has been widely rejected (see, e.g., Fodor et al., 1980).

The above work raises the question though, is such rule-like behaviour reflective of some 'natural' preference for rule-based cognitive processing in humans? Interestingly, Lea and Wills (2008) have recently challenged the view that unidimensional classification is a reliable sign of rule-based cognition. Their argument rests partly on the fact that single dimensions sometimes come to control the behaviour of nonhuman animals when these animals are presented with multidimensional stimuli. For example, in experiments that have employed artificial polymorphous concepts – that is, where category membership reflects the principle of family resemblance – analysis in birds has shown that different dimensions control behaviour to a different extent (i.e., one dimension was preferred; Lea & Harrison, 1978; Lea, Wills, & Ryan, 2006). This has also been found in the study of 'natural' concepts in nonhuman animals (e.g., discriminating between male and female faces; Troje, Huber, Loidolt, Aust, & Fieder, 1999). Moreover, Lea and Wills (2008) point out that pigeons learn discriminations faster when they are based on a single stimulus dimension, rather than when they are based on multiple stimulus dimensions. If one assumes, therefore, that unidimensional classification is rule-based, then one would have to conclude that pigeons, at least, sometimes elaborate rules (Lea & Wills, 2008). This, however, fits poorly with the majority of animal learning research. Of course, it is possible that unidimensional classification in humans *is* reflective of rule use, whereas in nonhuman animals it is reflective of, perhaps, limited attentional capacity, meaning that animals cannot process all the available stimulus dimensions at any one time (Lea & Wills, 2008; see also, Sutherland & Mackintosh, 1971). With respect to humans, however, one further possibility is that participants' bias for unidimensional classification in laboratory-based studies of human unsupervised categorisation is simply some artefact of the experiments themselves. This possibility is discussed further in the next section.

1.3.2 Reconciling the theoretical rejection of necessity and joint sufficiency and participants' preference for unidimensional unsupervised categorisation

To recapitulate, while on the one hand we have the theoretical rejection of the assumption of necessity and joint sufficiency in concepts (see Fodor et al., 1980), on the other hand there exists a large body of research documenting a strong bias for a '*classical-type*' classification strategy within laboratory-based unsupervised categorisation. In accounting for these contradictory results, an important starting point is the simple fact that, in the majority of unsupervised categorisation experiments, and indeed in the majority of categorisation studies *per se*, we are not dealing with naturalistic categories. Instead, participants are focused upon a limited number of highly structured, artificial stimuli, created most often from binary data sets (Malt, 1995). This fact becomes important when one considers the following point: with respect to category coherence, the classical view provides a strong account. That is, once a person has defined a basis for classification (e.g., all stimuli with four legs are members of category A and all stimuli with eight legs are members of category B), stimuli can be rapidly classified with confidence. This has a secondary benefit of enabling a high degree of *cognitive economy* within the account (even if this is at the expense of *informativeness*; Komatsu, 1992). Therefore, when engaged in a traditional laboratory-based unsupervised categorisation task – where participant motivation is generally low (Murphy, 2002) – participants will likely favour a classification strategy based on the principles of economy and coherence, rather than informativeness. It has also been suggested that other prevalent task constraints imposed in the literature, such as specifying the number of categories to be used (more often than not, two), may further encourage unidimensional classification.

This is because by restricting stimulus classification, the experiment will appear a lot more like a problem-solving task than a categorisation task. Consequently, participants will likely favour a readily verbalisable strategy when engaging in stimulus classification (Murphy, 2002). Moreover, as noted in Chapter 0, whereas laboratory-based unsupervised categorisation will often be intentional, everyday unsupervised categorisation will, for the most part, be incidental. As Love (2002) showed, this distinction produces meaningful differences in the nature of participants' unsupervised classification behaviour, with intentional categorisation encouraging more unidimensional classification. Further work by Love et al. (2004; to be discussed later in this chapter) has also shown that the abstract stimulus structure of Figure 1, introduced by Medin at al. (1987), should itself encourage unidimensional classification. This is particularly problematic because many researchers investigating unsupervised categorisation have employed this stimulus structure.

The upshot of all this is that participants' bias for unidimensional unsupervised categorisation, which has been documented in so many laboratory studies, is likely an artefact of the factors specified above. Of particular interest in this thesis is the likely influence of stimulus structure in biasing participants' classification behaviour (see Love et al., 2004). Specifically, if one stimulus structure is able to bias people towards unidimensional classification, then it makes sense that a different stimulus structure should be able to bias people towards multidimensional classification. This intriguing possibility is the subject of empirical investigation in Chapter 2 of this thesis. Furthermore, if the majority of natural unsupervised categorisation takes places incidentally (Love, 2002), then clearly unsupervised categorisation in the laboratory also needs to be assessed in an incidental manner. In Chapter 3 of this thesis, therefore, a new procedure is introduced to assess incidental unsupervised categorisation in the laboratory. By employing this procedure, I specifically sought to assess the factors that influence whether stimuli are incidentally classified together, or classified apart. It is important to note, however, that while much of the unsupervised categorization literature has documented a preference for unidimensional classification in humans, this is not to say that multidimensional unsupervised classification is *never* found.

1.3.3 Family resemblance in unsupervised categorisation

When discussing naturalness, the family resemblance view looks to the interplay between our perceptual system and the environment. That is, human category construction (and category coherence) is considered to reflect the natural discontinuities that exist in the environment, as perceived by the classifier (Komatsu, 1992). Unsupervised categorisation should, therefore, reflect a partitioning of a stimulus set in the manner most privileged by the human perceptual system (see Rosch, 1978; Rosch & Mervis, 1975). That is, unsupervised categorisation should maximise within-category similarity and minimise between-category similarity. Moreover, providing participants with the prototypical members of each category within a stimulus set should encourage classification based on overall similarity (or family resemblance).

As documented above, however, the majority of laboratory-based unsupervised categorisation research has shown a bias among people to engage in unidimensional classification. Furthermore, even when the prototypical members of the two experimenter-defined categories are presented, participants still seem to favour classification based on unidimensional rules (e.g., Medin et al., 1987). This, then, appears to argue against unsupervised category formation based on the principle of family resemblance. However, as noted in Section 1.3.2, the bias for unidimensional unsupervised classification in the laboratory is likely an artefact of the experiments themselves. So, do people ever spontaneously notice the family resemblance structure of an artificially created stimulus set? The answer is yes: using binary dimensioned stimuli, Billman and Knutson (1996) showed that participants were able to notice family resemblance structure when given a set of highly structured items (i.e., when many of the items' attributes covaried with each other). Given this finding, do people ever spontaneously categorise a set of items in concordance with a family resemblance principle? Again the answer is yes: Regehr and Brooks (1995), for example, found that while participants who engaged in a simultaneous categorisation task showed a preference for unidimensional classification, those that engaged in a sequential categorisation task produced more family resemblance sorting (cf. Milton & Wills, 2004).

Handel and Imai (1972), and, Kemler and Smith (1979) have further shown increased family resemblance sorting when participants are presented with integral as opposed to separable stimuli (cf. Milton & Wills, 2004). However, Kemler and Smith (1979) concluded that, in general, participants preferred single dimension sorts. Contrary to the findings of Milton et al. (2008), reported earlier, a number of studies have shown that family resemblance categorisation is promoted under conditions of increased time pressure. For example, using a minimal unsupervised categorisation paradigm, Ward (1983) found that those participants classed as 'slow responders' (whose median response latency was more than the group median) showed significantly fewer family resemblance sorts than those participants classed as 'fast responders' (whose median response latency was less than the group median). Similarly, Smith and Kemler Nelson (1984) found that those participants who engaged in a speeded unsupervised classification task produced significantly more family resemblance sorts than those participants that engaged in non-speeded classification. In comparing these findings to those of Milton et al. (2008), Milton et al. suggest a levels-of-time-pressure explanation. That is, the amount of time pressure imposed on classification shares a nonmonotonic relationship with the classification strategy imposed. Milton et al. (2008) demonstrate this by showing an increase in family resemblance sorting at stimulus presentation times of 256 ms and 640 ms, relative to presentation times of 64 ms and 384 ms. Finally, Smith and Kemler Nelson (1984) found that employing a concurrent cognitive load – here, having participants count backwards in 17s from a specified starting point – during classification also significantly increased family resemblance sorting to a level above that of unidimensional sorting (cf. Milton et al., 2008).

What is striking about these results, however, is that almost all of the unsupervised categorisation studies that have reported an increase in family resemblance sorting have only done so by introducing some additional manipulation. As highlighted by the findings of Ward (1983) and Smith and Kemler Nelson (1984), one of the most effective manipulations in this regard is speeded categorisation. Interestingly, the identification and categorisation of familiar everyday objects has been shown to occur very rapidly, at somewhere between 50-100 ms (see Grill-Spector & Kanwisher, 2005; Thorpe & Imbert, 1989). Does speeded categorisation in the laboratory most accurately reflect people's unsupervised classification behaviour in the real world, therefore? Well, perhaps; but, the identification and classification of novel objects will necessarily take longer. Moreover, at around 100 ms, Milton et al. (2008, Experiment 4) found that their participants clearly preferred unidimensional classification; indeed, family resemblance sorting was at one of its lowest levels. Also, despite showing increased levels of classification based on family resemblance, in general, participants in speeded categorisation studies have still shown an overall preference for unidimensional classification (but, see Smith & Kemler Nelson, 1984, Experiment 6, Concurrent task, Initial phase). What is apparent, therefore, is that the issue of unidimensional versus family resemblance unsupervised categorisation is not one that can be understood simply in terms of the speed of classification.

In conjunction with the reasons documented in Section 1.3.2, the lack of family resemblance sorting found in studies of unsupervised categorisation may further be attributed to two other factors: First, with the allowance of nondefinitional information within people's representations, it is arguable that the family resemblance view does not provide as strong an account of category coherence as the classical view. Consequently, for simple 'categorisation problems' presented in the laboratory, classification on the basis of family resemblance may be considered less cognitively efficient and economic (Komatsu, 1992). Second, when humans engage in real world categorisation, they will likely do so with the benefit of a great deal of associated background knowledge. As we will see in Section 1.3.5, the introduction of prior knowledge into unsupervised categorisation tasks can produce a marked increase in classification based on family resemblance. The question remains, however, if presented with a set of stimuli for which family resemblance classification is predicted to be 'most intuitive' on the basis of their abstract similarity structure, will people's categorisations reflect this prediction in the absence of any prior knowledge? As noted earlier, this question is the subject of investigation in Chapter 2 of this thesis.

1.3.4 Instances in unsupervised categorisation

To recapitulate, according to the instance approach of the exemplar view of conceptual structure, category membership is a function of the similarity of an encountered item to one or more of the instance representations that form a category (see, e.g., Nosofsky, 1986). The more similar an item is to a previously encountered instance, the more likely it is that that item will be categorised accordingly. Consequently, the more typical an item is of a particular category, the more likely it is that it will share a high degree of similarity with one or many of the stored instances forming that category, and so the more readily categorisation of that item will proceed. While some authors have argued that the instance approach offers no systematic explanation of category coherence (e.g., Komatsu, 1992), as has already been detailed, these arguments have been successfully countered on a number of fronts (e.g., Goldstone, 1994; Hahn & Chater, 1997; see Section 1.1.4.1). In light of this, as for the family resemblance view, coherent categories within the instance approach should reflect some optimal ratio between maximising within-group similarity and minimising between-group similarity (Rosch, 1975). As such, people's unsupervised categorisations should similarly reflect the principle of family resemblance (overall similarity).

As has already been shown, despite a number of studies documenting family resemblance sorting in laboratory-based unsupervised categorisation tasks, this has often only been achieved following the introduction of some other, critical manipulation (e.g., imposing a time constraint). This observation suggests, therefore, that classification based on a principle of family resemblance is not what participants regard as 'most intuitive' (at least with respect to the experimental tasks employed). Moreover, while these additional task manipulations have been found to increase the prevalence of family resemblance sorting, as noted above, unidimensional classification is often still preferred overall. Is there any specific evidence for unsupervised categorisation based on an instance-based principle? Using simple butterfly-like stimuli, Milton and Wills (2004) recently reported that participants who were presented with a spatially separable form of their stimuli (i.e., the antennae, wings, etc., were presented separately, but next to each other) showed significantly increased family resemblance sorting compared to those participants who were presented with a spatially integrated form of their stimuli. This surprising result was confirmed using different, lamp-like stimuli (see Milton & Wills, 2004). Milton and Wills (2004) provide an explanation for their findings by proposing that people use an analytic, *dimensional summation* strategy in categorization.

Milton and Wills (2004) argue that, when categorising by family resemblance, participants engage in a process whereby they individually focus on each dimension of variation, and then categorise a stimulus on the basis of whether it has more characteristic features of, for example, Category A members or Category B members. Critically, therefore, this analytic dimensional summation process appears to argue against the idea that novel stimuli are compared to some stored, averaged abstraction of a category as a whole (as is suggested by the prototype view of conceptual structure). Moreover, the dimensional summation hypothesis encourages the view that unidimensional and family resemblance classification strategies are based on similar cognitive processes (Milton & Wills, 2004). Indeed, this idea receives some support from demonstrations that, when appropriately weighted, exemplar models can readily account for both unidimensional and family resemblance categorisation (see Nosofsky & Johansen, 2000). Based on this view, then, classification based on family resemblance is simply a more sophisticated version of dimensional summation than is unidimensional classification. Consequently, dimensional summation in categorisation provides a ready explanation for participants' preference for unidimensional classification in the laboratory: that is, it is cognitively less effortful than family resemblance sorting.

However, the dimensional summation hypothesis appears odd for a couple of reasons. First, a number of studies have reported the opposite result to Milton and Wills (2004), finding that spatially integrated stimuli are more likely to encourage family resemblance sorting (e.g., Garner, 1974). Second, if dimensional summation is the mechanism for categorisation, then why does natural categorisation reflect a principle of family resemblance (e.g., Rosch & Mervis, 1975; Rosch et al., 1976)? That is, we do not normally view stimuli in the environment broken up into their constituent parts. Based on the dimensional summation hypothesis, therefore, humans' everyday concepts and categories should be biased towards unidimensional classification. One could argue, of course, that given the amount of experience that humans have with everyday stimuli, this allows the supposed more effortful family resemblance classification strategy to proceed successfully. However, this seems somewhat odd in light of findings by Smith and Kemler (1977), who showed that children are more likely to engage in family resemblance sorting than are adults. If family resemblance classification is a more complex analytic process, surely Smith and Kemler (1977) should have found the reverse result. Moreover, support for the view that overall-similarity-based classification is a more associationistic process stems from numerous studies showing that nonhuman animals readily engage in classification based on family resemblance (Lea & Wills, 2008). For example, experiments that have used 'natural concepts', such as HUMAN, TREE, or FISH, have shown that pigeons can readily learn to discriminate between different complex scenes based on the presence or absence of these 'concepts' (e.g., Cerella, 1979; Herrnstein & Loveland, 1964; Herrnstein, Loveland, & Cable, 1976; Siegel & Honig, 1970; Wasserman et al., 1988). Furthermore, nonhuman animals can also generalise their initial discrimination learning to novel pictorial scenes. The reason why such discrimination learning is considered to be based on overall similarity is because these 'natural concepts' are considered to be polymorphous; that is, deciding category membership is assumed not to be possible on the basis of a singly necessary or sufficient feature (Herrnstein, 1985; Jitsumori, 1993).

However, given the complexity of these 'natural concept' scenes, it is rather difficult to know exactly how nonhuman animals are solving these kinds of discriminations (see D'Amato & Van Sant, 1988; Troje et al., 1999; but, see Jitsumori, 1993, 1994). While these studies do appear to show that nonhuman animals can, in principle, discriminate on the basis of overall similarity, this does not mean that this form of discrimination learning is *necessarily* most natural to them (see Lea & Wills, 2008). Importantly, these studies also do not definitively document categorisation (in any meaningful sense) in nonhuman animals (see Chater & Heyes, 1994). Before one can ask whether nonhuman animals have a natural preference for categorisation based on a principle of family resemblance, therefore, one first needs to address the question of whether nonhuman animals engage in unsupervised (spontaneous) categorisation at all. Consequently, although the issue of unidimensional versus multidimensional categorisation in nonhuman animals is not directly investigated in this thesis, a new method for investigating unsupervised (spontaneous) categorisation in nonhuman animals is introduced in Chapter 3.

In summary, the above studies document an overall bias for unidimensional classification in laboratory assessments of human unsupervised categorization. Indeed, even when increased levels of family resemblance sorting have been shown, often unidimensional classification is still preferred overall. One obvious difference between laboratory-based and real world categorisation, however, is that everyday categorisation takes place with respect to a person's understanding about the world, and the objects that exist within it. Consequently, perhaps it is this theoretical knowledge that predominantly encourages classification based on family resemblance (see Murphy & Medin, 1985).

1.3.5 The influence of theoretical knowledge on unsupervised categorisation

If theoretical knowledge does indeed form the basis for category coherence, then one would naturally expect this to be reflected in people's categorisation behaviour. What would this mean with respect to laboratory-based unsupervised categorisation? If, as is widely accepted within cognitive psychology, real world category structures are rich and multidimensional, then a clear prediction would be that incorporating theoretical knowledge into an artificial unsupervised categorisation task should encourage participants to engage in classification based on family resemblance. A number of authors have demonstrated how prior knowledge about a set of stimuli influences participants' unsupervised classification behaviour. For example, Lassaline and Murphy (1996) found that participants who had previously been asked inductive questions about verbal and pictorial stimuli engaged in significantly more family resemblance sorting in a subsequent unsupervised categorisation task than participants who had simply been asked 'frequency questions' about the stimuli, or who simply completed the sorting task. An induction question took the form, "If *X* has the property *Y*, what kind of *Z* does it have?" (*X* may refer to an animal, *Y* to tail length (long or short), and *Z* to tooth shape (flat or sharp)). In contrast, a frequency question focused only on single dimensions, such as, "How many animals have a short tail?" (Lassaline & Murphy, 1996). Similarly, Spalding and Murphy (1996) showed that people are able to spontaneously utilise their background knowledge when afforded to do so in laboratory-based unsupervised categorisation tasks. For example, participants who were able to thematically relate the features of a set of stimuli produced significantly more classifications based on family resemblance than participants who were unable to thematically relate the features of a set of stimuli. Furthermore, Kaplan and Murphy (1999) found that even if only a single feature per item is associated with prior knowledge, then this will still help participants notice the family resemblance structure for a set of stimuli.

Ahn (1990, 1991) has also shown the importance of background knowledge in influencing participants' production of family resemblance sorts. In her experiments, participants were either presented with the prototypical members of each category to use as templates for category construction (the prototype condition), presented with information that identified certain features (e.g., a flower being brightly coloured) with some particular property (e.g., being particularly attractive to birds rather than bees; the theory condition), or were given no additional information (the control condition). In a subsequent unsupervised categorisation task, she found that while those participants who had received no additional information showed an overwhelming bias for unidimensional classification, those participants in the other two conditions tended to use a family resemblance principle as the basis for their unsupervised categorisations. Notably, however, those participants in the theory condition were more likely to engage in family resemblance-based unsupervised classification than participants in the prototype condition.

The experiments described in this section demonstrate how the introduction of prior knowledge into unsupervised categorisation tasks increases the prevalence of family resemblance sorting. However, in line with other findings, where family resemblance sorting has been increased by prior knowledge, participants often still show an overall preference for unidimensional unsupervised categorisation. While these findings provide some support for the view that prior knowledge plays an important role in guiding human categorisation (Murphy & Medin, 1985), it is also apparent that prior knowledge does not uniquely determine family resemblance-based classification (at least in the laboratory). Critically, prior knowledge appears to play an important role in highlighting (enhancing) the correlated nature of stimulus attributes, but this can only guide unsupervised classification so far. It appears, therefore, that what is most critical in determining humans' overall bias for unidimensional or multidimensional unsupervised classification is the nature of the underlying stimulus structure (i.e., the similarity-based relations that exist between a set of stimuli). That is, if the similarity structure of a set of stimuli is biased towards unidimensional categorisation (i.e., unidimensional classification is 'more intuitive'), then participants' classifications will likely reflect this bias (see Figure 1; Love et al., 2004). To reiterate, Chapter 2 of this thesis sought to directly test the influence of stimulus similarity structure on the issue of unidimensional versus multidimensional unsupervised classification.

1.3.5 Overview of unsupervised categorisation research: what does it tell us about human conceptual structure?

The laboratory-based investigation of human unsupervised categorisation has revealed some surprising results. Despite the theoretical rejection of the classical view of conceptual structure (that is, category formation based on necessary and jointly sufficient features), many unsupervised categorisation studies have documented a robust and overwhelming bias for a classical-type, unidimensional classification strategy among people. Put simply, this laboratory-based preference does not fit with current understanding about the nature of our everyday category structures, which are patently based on a principle of family resemblance (Rosch & Mervis, 1975; Wittgenstein, 1953). This is not to say that unsupervised classification based on a family resemblance principle is *never* found in the laboratory. Manipulations of stimulus format, procedure, etc., have all influenced the issue of unidimensional versus family resemblance sorting. The fact that some additional task manipulation has been required to increase family resemblance sorting, however, suggests that this kind of unsupervised classification is not participants' preferred strategy for categorisation. Moreover, where family resemblance sorting has been increased, more often than not an overall preference for unidimensional unsupervised categorisation has remained.

What, then, does this research tell us about the nature of human conceptual structure? First, the evidence suggests that the human cognitive system does not spontaneously recognise and utilise the family resemblance structure of a set of items (but, see Billman & Knutson, 1996). Rather, when possible, it appears much simpler to group a set of items on the basis of some necessary feature along a single dimension of variation. So, is the use of necessary and sufficient features in human categorisation natural? Well, perhaps for categorisation tasks undertaken in unnatural situations using artificially created category structures. However, what is also clear is that, when provided with only a minimal amount of prior knowledge that causally relates features within a stimulus set, participants produce more naturalistic, family resemblance sorts (e.g., Kaplan & Murphy, 1999, 2000; Lassaline & Murphy, 1996; Spalding & Murphy, 1996). Do 'naïve theories', acquired from prior knowledge, underlie category coherence, therefore? The most likely answer to this question is that prior knowledge commands an important influence over human categorisation, but it is unlikely to be the sole, or even the main, determinant of stimulus classification. Similarity, ever ready to impress itself, is clearly a critical component in guiding and determining human (and nonhuman animal) categorisation. Furthermore, many of the arguments that cast doubt over the suitably of similarity as providing a basis for human categorisation have now been addressed (see Goldstone, 1994; Hahn & Chater, 1997; Medin, et al., 1993). These authors have convincingly argued that similarity is not too unconstrained to afford the constrained nature of categorisation. While prior knowledge may be central in highlighting the interconnected nature of stimulus attributes, this does not mean that it is the 'glue' that holds categories together (Murphy, 2002).

As documented in Section 1.3.2, a number of factors have likely played an important role in producing the overwhelming bias for unidimensional unsupervised classification in the laboratory (e.g., constrained categorisation, etc.). To my mind, one of the most important of these factors is the fact that the stimulus structure of Figure 1 has been repeatedly employed in studies of human unsupervised categorisation. This is important because, although Medin et al. (1987) assumed that this stimulus structure would naturally promote family resemblance sorting, modelling work has predicted that this structure should actually be considered 'most intuitive' in terms of classification based along a single dimension of variation (see Ahn & Medin, 1992; Love et al., 2004). Given this fact, it is therefore imperative to determine the influence of stimulus similarity structure on the issue of unidimensional versus multidimensional unsupervised classification. That is, an assessment of unsupervised categorisation needs to be made where classification is entirely unconstrained, and where modelling work has established the similarity-based biases that exist within the stimulus structure(s) being used. Specifically, based solely on abstract similarity structure, one needs to contrast a situation where a preference for unidimensional unsupervised classification is predicted with a situation where a preference for family resemblance-based unsupervised classification is predicted. Only if unidimensional categorisation persists in both these conditions can one start to draw firm conclusions about the 'naturalness' of this classification behaviour. These fundamental issues are the subject of experimental investigation in Chapter 2 of this thesis. To achieve this goal, detailed modelling work of human unsupervised categorisation is required. Fortunately, as noted in Chapter 0, there now exist a number of influential models of unsupervised categorisation within the psychological domain, and it is to these that I now turn my attention.

1.4 Modelling Advances in Unsupervised Categorisation

Despite evidence supporting the principles of necessity and joint sufficiency as a basis for human unsupervised categorisation, modelling work on this topic has traditionally taken similarity as its starting point (e.g., Fisher, 1996; but see, Ahn &

Medin, 1992). The modelling effort for unsupervised categorisation has largely been overshadowed by that on supervised categorisation; the latter producing some of the most influential modelling approaches within cognitive psychology, such as exemplar (Kruschke, 1992; Nosofsky, 1986, 1989) and prototype theory (Hampton, 2003). However, a plethora of models of unsupervised categorisation now exist from work done within psychology, statistics, and machine learning. All these three areas of research share the same problem of how to divide up large amounts of information in the 'best' way possible; of course, work in machine learning and psychology has a specific emphasis on mirroring, or modelling, what humans do naturally. An influential early account of unsupervised learning was proposed by Fried and Holyoak (1984). For Fried and Holyoak's account to be successful, however, knowledge of the number of categories sought by the category learner must be known *a priori* (see also, *K*-means clustering, Banfield & Bassill, 1977; Kohonan neural network architecture, e.g., Schyns, 1991), and the category density functions that the approach relies on must have a specific form (Pothos & Chater, 2002). Given the clear limitations of this approach (e.g., in the real world people do not generally know how many categories they should construct), it is not surprising that numerous models now exist that require no knowledge of the number of categories sought – although most still rely on the data conforming to a specific form, albeit in a much more flexible manner (e.g., AutoClass, Cheeseman & Stutz, 1995; CODE, Compton & Logan, 1993, 1999; COBWEB, Fisher, 1987, 1996; Fisher & Langley, 1990; see also, Corter & Gluck, 1992). An early model of unsupervised categorisation stemming directly from psychological research is Ahn and Medin's (1992) two-stage model of category construction. This model was developed in response to the robust finding of a strong bias among participants for unidimensional unsupervised classification (e.g., Medin et al., 1987). Briefly, classification is initially sought on the basis of a single dimension (Ahn & Medin, 1992, viewed unidimensional classification as a ubiquitous feature of human unsupervised categorisation). If a suitable classification cannot be identified on the basis of any single dimension, then a grouping based on a principle of family resemblance is subsequently identified. With regard to the number of categories sought by the two-stage model of category construction, the model predicts that this will mirror the number of values (reflecting individual features) along the dimension that is regarded as most salient (Ahn & Medin, 1992).

With these models in mind, I now focus on three models of unsupervised categorisation that have established themselves within the cognitive psychology literature. These three models are the Rational model (Anderson, 1991), the simplicity model (Pothos & Chater, 2002), and SUSTAIN (Love et al., 2004). While it is an interesting task in and of itself to review these models to better understand how they implement unsupervised categorisation, these three models will also be considered in the experimental work of Chapter 2 of this thesis. In particular, the simplicity model of unsupervised categorisation was employed to generate the predictions for the experimental work presented in Chapter 2. Therefore, a somewhat more detailed description of the simplicity model is presented below.

1.4.1 The Rational Model

Anderson's (1991) Rational Model is a Bayesian model of human categorisation, viewing human categorisation as a product of its necessary adaptation to the environment (Anderson, 1991). While the role of similarity is not central in the Rational Model, categorisation predictions often reflect the overall similarity structure of the stimulus domain. Indeed, as Anderson states, "The probability of an item coming from a category is a function of its feature similarity" (1991, p. 415). In the context of categorisation, there are two key considerations that the Rational Model has to address: First, what property of categorisation is the human cognitive system trying to optimise; second, what is the structure of the environment in which the human cognitive system has evolved, and as such, is adapted to? This strong emphasis on the environment is reflective of the view of Rosch and her colleagues (Rosch & Mervis, 1975; Rosch et al., 1976), and given the probabilistic nature of the environment, Anderson suggests that humans "start out with some weak assumptions about the environment and with experience make these increasingly strong" (1991, p. 409).

The Rational Model is an iterative (or incremental) model of learning and categorisation. From a starting position where no categories are specified, at each step, it decides how a novel instance should be categorised. In this way, it slowly builds a classification for a set of stimuli. A key constraint on the Rational Model's incremental strategy of category formation, however, is that it does not consider all possible category structures for a given data set (this being due to the high computational demands that such a procedure would require). Instead, the model hypothesises some specific category structure of the objects seen, and then the algorithm commits to this. When a new item is presented, this hypothesised category structure may have to be altered from its previous state, and so the model commits to a hypothesis about the category structure of the objects after every object seen (see Anderson, 1991, for a more in-depth discussion of this constraint). Critical to Anderson's (1991) theorising is the premise that humans need to have well-formed category structures at their disposal at all times; these naturally being updated after every new object is presented. The incremental nature of the Rational Model means that the model gives rise to order effects.

When a new object is presented, the Rational Model calculates for each category *k* the probability $P(k \mid F)$ that the new object belongs to category *k* given that the new object has features *F*. $P(k \mid F)$ is calculated from two terms, $P(k)$ and $P(F \mid k)$. The first of these terms is a prior probability; that is, before the feature structure of an object has been assessed, the prior probability of that object coming from category *k*. The second term is a conditional probability; that is, the probability of having features *F* given that it comes from category *k*. So, a new instance with feature structure *F* is classified to the category *k* for which the product $P(k)P(F \mid k)$ is greatest (or, it may be assigned to a new category). For example, if you see a new object that looks like a 'cat', assign it to the category of cats, since the feature structure of the object is most probable given this category membership.

Surveying the literature, Anderson (1991) showed that the Rational Model provided an accurate description of a number of important experimental findings. For example, in a test of categorisation to a face prototype, the predictions of category membership of 25 test stimuli made by the Rational Model correlated extremely well (.90) with participant categorisations from the original study by Reed (1972). Similarly, the Rational Model was shown to be sensitive to the frequency of presentation of certain exemplars, as were participants in Nosofsky (1988b). It also accurately predicted when linearly nonseparable categories (that is, when it is not possible to draw a straight hyperplane in the category space that separates a set number of categories) would be easier to learn than linearly separable categories (that is, when it is possible to draw a straight hyperplane that separates a set number of categories; see Medin & Schwanenflugel, 1981). Anderson (1991) further showed that, in the absence of any feedback, the Rational Model was able to identify category structure within the materials used by Homa and Cultice (1984). Furthermore, the

Rational Model has been successfully applied to basic level categorisation, simulating results by Murphy and Smith (1982; see, Anderson, 1990).

In summary, the Rational Model of human categorisation has been shown to be an effective model of supervised categorisation that also extends into the identification of category structure when no feedback is available. Anderson makes a number of conclusions about his Rational Model, two of which are particularly interesting in the current context: First, he proposes that "a good case has been made for the proposition that categorisation behaviour can be predicted from the structure of the environment at least as well as it can from the structure of the mind" (1991, p. 427). Second, after acknowledging that the Rational Model has little to say with specific regard to what the structure of the mind is, he suggests "that the mind has the structure it has because the world has the structure it has" (1991, p.428).

1.4.2 The simplicity model of unsupervised categorisation

The second model to be discussed, the simplicity model of unsupervised categorisation (Pothos & Chater, 2002), provides a computational formalism for Rosch and Mervis's (1975) proposal that 'good' categories are ones that maximise within-category similarity and minimise between-category similarity. Consequently, the simplicity model predicts what has become regarded as people's preferred level of categorisation; that is, the basic level of categorisation (as opposed to superordinate or subordinate level categorisation; Rosch et al., 1976). A number of clustering algorithms have been proposed with the goal of trying to extract the 'natural' or 'best' way of partitioning a set of items (that is, through forming 'good' or 'intuitive' categories; see Krzanowski & Marriott, 1995). This contrasts with hierarchical clustering models where one typically ends up with a single cluster, unless a 'cut-off' criterion is specified. The critical issue with regards to these clustering algorithms is what determines 'bestness'. In all clustering methods, partitioning of a set of items is deemed to reflect regularity in the similarity structure of those items (Pothos & Chater, 2001); but, how does one measure how good a specific classification is? For example, one may identify two plausible ways of partitioning an item set based on their similarity structure, but which classification (partitioning) is to be preferred? As an answer to this problem, Pothos and Chater (2002) propose the notion of simplicity in the form of the minimum description length principle (MDL; Rissanen, 1978). The MDL principle reflects the idea that shorter descriptions of a given data set (specified in terms of a codelength in some (universal) programming language) are better. That is, the shorter the codelength needed to describe the data, and therefore the data itself, the better that description (or 'theory') is (Pothos & Chater, 2001; see also, Quinlan & Rivest, 1989). To specify a codelength for the similarity structure of a set of items, Pothos and Chater (2001, 2002) take their reference from information theory (this will be discussed further shortly).

For the purpose of the simplicity model, 'descriptions' can be mapped directly onto classifications, and as such, a codelength can be associated with a specific classification. Following from Rissanen (1978), Pothos and Chater propose that "According to the simplicity model...groupings associated with a short codelength (high compression) will be favoured" (2002, p. 310). Initially, therefore, the simplicity model computes, in bits, the codelength required to describe the total similarity information in a set of items without any categories (i.e., the *raw* similarity information). Certain clustering patterns will reduce the codelength required to describe the similarity information of the item set more than others; these clustering patterns will therefore be preferred, given that they will achieve an overall greater compression of the raw description of all the similarity information. Specifically, in line with the proposal of Rosch and Mervis (1975), the simplicity model looks to achieve a clustering pattern whereby the similarity of items within a cluster (a collection of items in a set) is greater than the similarity of items between clusters (Pothos & Chater, 2001, 2002).

To avoid becoming embroiled in the debate over the nature of similarity (see, e.g., Goodman, 1972; Hahn & Chater, 1997; Medin et al., 1993; Tversky, 1977), the simplicity model looks to capture the broadest view of similarity information possible. That is, given four items *A*, *B*, *C*, and *D*, for example, the simplicity model asks, is similarity (*A*, *B*) less or greater than similarity (*C*, *D*), without any regard to how similarity is defined. Judgements of *pairwise inequalities* reflect a binary decision: for example, similarity (*A*, *B*) is either greater than similarity (*C*, *D*), or it is less than similarity (*C*, *D*; Pothos & Chater, 2001). Pothos and Chater (2001) note that ties (or *equalities*) can also be accounted for within these judgements; however, they propose that ties are extremely unlikely with real-valued domains, and so they are ignored for simplicity. Of course, while this may be true for real-valued domains, ties will likely occur with the kinds of materials that most unsupervised categorisation studies employ; namely, stimuli constructed from binary-valued feature dimensions. In practice, therefore, the simplicity formalism is slightly adjusted to take into account equalities. One further assumption made in the simplicity model is that similarities obey the metric axioms of symmetry and minimality. That is, similarity (*A*, *B*) = similarity (*B*, *A*), and similarity (*A*, *A*) = maximum similarity. These metric axioms are assumed to be obeyed when meaningless, schematic stimuli are employed (the introduction of general knowledge factors into classification judgements will, however, likely lead to violations of these metric axioms; see Tversky, 1977).

So, as stated above, the simplicity model initially computes (in bits) the codelength required to describe the total similarity information in a set of items without any categories (i.e., the *raw* similarity information). For example, for 10 objects, there are 10×(10-1)/2 = 45 unique similarities ('unique' here represents the fact that if similarity (*A*, *B*) is included, then similarity (*B*, *A*) will not be included, and that similarity (*A*, *A*), etc., will also not be included). Consequently, there are 45×(45-1)/2 = 990 unique pairs of similarities ('unique' here represents the fact that if the relation similarity (*A*, *B*) > similarity (*A*, *C*) is included, then the relation similarity (*A*, *C*) < similarity (*A*, *B*) will not be included, and that relations like similarity (*A*, *B*) > similarity (*A*, *B*) will not be included). Overall, therefore, for 10 objects there are 990 pairs of similarity relations (inequalities), meaning that a codelength of 990 bits is required to describe the corresponding similarity information.

In the simplicity model, Pothos and Chater (2002) *defined* categories as imposing *constraints* on the similarity relations between pairs of stimuli. To recapitulate, the definition Pothos and Chater (2002) used was that all similarities within categories are assumed to be greater than all similarities between categories (Rosch, 1975). So, if one assumes, for example, that the 10 objects specified above can be clustered into two perfect categories, with five objects in each category (i.e., no constraints are violated), then there are 5×(5-1)/2 = 10 within-category similarities. In total, therefore, there are 20 within-category similarities when considering both categories together. Moreover, there are 5 × 5 between-category similarities. Consequently, given these two perfect categories, there are a total of 20 × 25 = 500 constraints. By imposing categories, then, the codelength required to describe the similarity structure of the objects is now, approximately, 990-500 = 490 bits. "Approximately" here reflects two points: First, the simplicity model also needs to take into account the codelength required to select the 'best' classification from all possible classifications of *r* items; that is, the complexity of specifying the category membership of a set of items (see Pothos & Chater, 2001, 2002). This is done using Stirling's number, $\sum_{v=0}^{n}(-1)^{v}\frac{(n-v)^{r}}{(n-v)!v!}$, which describes the number of ways r items can be divided into n categories. Second, when a particular classification is imposed, in general, some of the constraints will be wrong. These wrong constraints need to be corrected, therefore, so as to reconstruct the data (Pothos & Chater, 2002). Pothos and Chater (2002) detail that if there are u constraints, of which e are erroneous, then the total code for correcting erroneous constraints is $\log_2(u+1)+\log_2\left({}_{u}C_{e}\right)$ bits, given that there are ${}_{u}C_{e} = \frac{u!}{e!(u-e)!}$ ways to choose e items from a set of u.

In summary, the simplicity model provides a metric for assessing 'category intuitiveness', by determining how much simpler the description of a stimulus structure is *with categories*, compared to *without categories*. Overall, the 'goodness' of a classification will be better the more constraints and fewer errors there are; this will be reflected in a corresponding reduction in description length. In general, simplicity model predictions are typically specified as the ratio of codelength (with categories) / codelength (without categories), expressed as a percentage. Therefore, the lower this percentage, the greater the 'simplification' of the code achieved by imposing a classification and the more psychologically intuitive (obvious) the classification is predicted to be. For brevity, this percentage is referred to as 'codelength'. Classification codelengths typically vary between 50% and 100% (as said, lower values indicate a more psychologically intuitive classification). The computation of the different codelength terms specified above is effectively an application of the formal simplicity framework of Minimum Description Length (Rissanen, 1989). The simplicity model is run in a straightforward way: its input is the coordinates of a set of stimuli when represented in an assumed psychological space, out of which the model generates information about pairs of similarities (typically using the Euclidean metric). The model employs a search algorithm to identify the best possible classification for the set of items. The algorithm is akin to agglomerative clustering ones, which initially assume that all items belong to separate categories, and then gradually combine items to try to improve this classification. Unlike many prominent models of categorisation (whether they model supervised or unsupervised categorisation), the simplicity model is parameter free (Pothos & Chater, 2002).

Over a series of four experiments, Pothos and Chater (2002) presented experimental support for their simplicity model of unsupervised categorisation. These experiments included having participants simply draw lines around a set of data points in a manner that they felt represented the most natural and intuitive partitioning for those data points; grouping together sets of star stimuli, which were constructed from the coordinates of a specified stimulus structure; and, providing pairwise similarity ratings for 11 black and white polka dot square stimuli. In conclusion, the simplicity model provides an important and interesting metric for computing the 'intuitiveness' of a classification. Moreover, one is able to determine an 'optimal' classification for a set of stimuli without the need for any free parameters. The development of the simplicity model adds to the growing literature in which the simplicity principle has been applied to explain a number of cognitive processes (see Chater, 1999; also, e.g., Hahn et al., 2003).

1.4.3 SUSTAIN

Like the Rational Model (Anderson, 1991) and the simplicity model (Pothos & Chater, 2002), the Supervised and Unsupervised STratified Adaptive Incremental Network (SUSTAIN; Love et al., 2004) assumes that the world has some natural structure that the human perceptual and conceptual systems exploit (e.g., Rosch & Mervis, 1975). At the heart of SUSTAIN's categorisation behaviour, however, is the flexible search for structure. Indeed, Love et al. (2004) note that the most intuitive structure for a set of items based solely on perceptual similarity may not always be as useful as the structure derived from an alternative analysis. The promotion of flexibility in search by Love et al. is highlighted by the following passage: "Thus, the categorisation system must be able to both assimilate structure and discover or even create that structure" (2004, p. 309). This broad notion of categorisation fits nicely with Malt's (1995) conclusion that a structured environment in itself is insufficient to determine categorisation, although it clearly plays a key role in categorisation (see also, Mervis & Rosch, 1981; Wisniewski & Medin, 1994).

SUSTAIN is particularly interesting as it tries to capture the full continuum of human categorisation behaviour, from unsupervised categorisation to supervised categorisation, in a single model. In contrast to some models (e.g., backpropagation models), SUSTAIN has an adaptive architecture to learning (Love et al., 2004). That is, it initially searches for simple solutions to a particular categorisation problem and only expands the complexity of such solutions when the problem requires. Like the Rational Model, SUSTAIN is an incremental model of category learning, and as such, is susceptible to ordering effects (Love et al., 2004; see, e.g., Bruner, Goodnow, & Austin, 1956). With specific regard to SUSTAIN's modelling of unsupervised categorisation, once again similarity has a central role to play in initially determining structure within a given stimulus set, in line with both the Rational Model and the simplicity model. The intuition is that similar items will tend to cluster together, favouring groupings that maximise within-category similarity and minimise between-category similarity (Rosch & Mervis, 1975). Moreover, SUSTAIN's unsupervised categorisation component is also driven by the fact that it reacts to 'surprising' events. That is, if a novel item is encountered that does not fit well into any existing clusters (i.e., the similarity between an item and the cluster the item is most similar to is below a certain threshold, Love et al., 2004) then a new cluster will likely be created.

SUSTAIN is composed of the following basic components: a set of input units; a set of clusters that compete to respond to an input stimulus; a set of output units that mirror the input layer and serve as the corresponding inputs to a decision procedure; the decision procedure that generates a response (see Love et al., 2004). Stimuli are represented in terms of vector frames, "where the dimensionality of the vector is equal to the dimensionality of the stimuli" (Love et al., 2004, p. 313). Along with the perceptual dimensions (e.g., colour), these vector frames also include the category label as a stimulus dimension. Similarity is a function of the distance between vector frames within a multidimensional representational space; the smaller the distance between two vector frames, the more similar those items are taken to be. In the model simulations of Love et al. (2004), they only focused on stimuli whose dimensions are nominal, rather than continuous. While Love et al. (2004) note that SUSTAIN can represent continuous-valued stimulus dimensions, the details on how this is done are somewhat sketchy. However, to represent multiple-valued, nominal stimulus dimensions, at least, SUSTAIN simply recruits multiple input units (Love et al., 2004). With respect to unsupervised categorisation, an important free parameter in SUSTAIN is its cluster recruitment mechanism; that is, the mechanism that specifies the threshold of dissimilarity required between a novel item and an already formed cluster for the novel item to create a new cluster, rather than become part of an existing cluster. This parameter is important as it indirectly determines the number of categories created. However, despite representing a free parameter (having a range between 0 and 1), for simplification of analysis, this parameter was arbitrarily fixed at .5 for all simulations run by Love et al. (2004).

With respect to previous unsupervised categorisation research, Love et al. (2004) showed that SUSTAIN accurately predicts the contradictory findings of Billman and Knutson (1996) and Medin et al. (1987). To recapitulate, Billman and Knutson (1996) found that participants who received a set of highly structured stimuli, in which there were many features intercorrelations, performed better in a later classification task than participants who received a set of poorly structured stimuli, in which the stimulus features were nonintercorrelated. Specifically, participants in the intercorrelated structure condition became aware of the family resemblance structure of the stimuli. In contrast, when engaged in unsupervised classification, Medin et al. (1987) found that their participants preferred unidimensional classification, even when the stimuli's feature dimensions were intercorrelated. On the basis of Billman and Knutson's (1996) results, this should have led to classification based on a principle family resemblance. SUSTAIN successfully reconciles these seemingly contradictory patterns of results in two ways: First, by focusing on the statistical regularities that existed within the category structures used by the authors (note, for example, that in contrast to the materials used by Medin et al., 1987, perfect correlations existed between the stimulus dimensions used in Billman & Knutson's, 1996, study). Second, by the fact that SUSTAIN is biased to focus on a small subset of stimulus dimensions when considered acceptable (Love et al., 2004).

In conclusion, the interplay between the unsupervised and supervised categorisation components of SUSTAIN makes it an extremely flexible model, and one that is more powerful than either a pure model of unsupervised or supervised categorisation. However, as will be discussed in Chapter 2, the existence of SUSTAIN's recruitment mechanism as a free parameter can result in ambiguity in its predictions over a number of important issues within unsupervised categorisation research.

1.4.4 Overview of modelling of unsupervised categorisation

A number of commonalities exist between the three models reviewed above: first, they share the assumption that humans perceive an environment that has a structured nature; second, they assume that humans are sensitive to these perceived structural regularities; third, they all invoke a central role for similarity in unsupervised categorisation. Essentially, the three models look to principles that seem at odds with many of the findings from investigations of laboratory-based unsupervised categorisation. All the models suppose that humans will construct categories in a manner that is consistent with the most easily identifiable structure within a given stimulus set. Most notably, in contrast to past experimenters' intuitions, Love et al. (2004) have shown that SUSTAIN *predicts* that unidimensional classification of the binary stimuli employed by Medin et al. (1987; see Figure 1) should be preferred, on the basis of their abstract similarity structure. Given the prevalence of this stimulus structure (see Figure 1) within the unsupervised categorisation literature, SUSTAIN's prediction clearly casts doubt over the reliability of any conclusion that participants are 'naturally' biased towards unidimensional classification. Rather, this prediction supports the view that the prevalence of unidimensional unsupervised classification in the laboratory is most likely an artefact of the similarity structure of the stimuli employed.

1.5 Summary and conclusions

Research on human unsupervised categorisation has documented an overwhelming bias for unidimensional classification, which appears to reflect a 'classical-type' view of categorisation. The classical view of human conceptual structure has, however, been widely discredited (see Fodor et al., 1980). In its place, theories of conceptual structure have been developed that emphasise the role of similarity in human categorisation (i.e., prototype and exemplar views). These latter theories have led to the development of influential models of human categorisation (e.g., Hampton, 2003; Nosofsky, 1986, 1988a, 1988b, 1989; Nosofsky et al., 1989, see also Kruschke, 1992), which have proved extremely successful in modelling a range of classification data. With regard to the modelling of unsupervised categorisation, a number of influential models have been outlined in this chapter, which again emphasise the role of similarity in classification. Essentially, these models assume that while similar stimuli should be 'spontaneously' classified into the same category, dissimilar stimuli should be 'spontaneously' classified into different categories. With respect to the Rational model, the simplicity model, and SUSTAIN, unsupervised categorisation is guided by the assumption that humans are sensitive to perceived regularities in the environment. However, SUSTAIN (Love et al., 2004) also incorporates a role for the human classifier, with stimulus classification being influenced by the goals of the classifier at the time of categorisation.

As noted in Chapter 0, and highlighted at the beginning of this chapter, the respective roles of a structured environment and the human classifier in determining 'spontaneous' category construction is a particularly interesting topic of discussion. If one were to conclude, for example, that the theory theory provided an accurate representation of human categorisation, then this would have critical implications for comparative assessments of categorisation (i.e., comparing human categorisation behaviour to that of nonhuman animal categorisation behaviour). That is, the higher-level cognitive account of the theory theory view means that it must deny the possibility of nonhuman categorisation based on the same underlying principles as human categorisation. Consequently, human and nonhuman animal categorisation will necessarily be qualitatively different (see Chater & Heyes, 1994). To reiterate, this account of human conceptual structure clearly suggests a view of categorisation in which the role of the classifier dominates. However, if human classification is predominantly influenced by the statistical properties of the environment (in terms of perceived structural regularities), then this at least allows the possibility that nonhuman animal categorisation is qualitatively similar to that of human categorisation.

With these issues in mind, this thesis seeks to better understand how stimulus similarity structure, and the statistical properties of the environment, guide and influence categorisation behaviour in humans and rats. Specifically, in the next chapter of this thesis (Chapter 2) I investigate the influence of abstract stimulus structure on the issue of unidimensional versus multidimensional unsupervised classification in humans. As I have argued in this chapter, one likely factor in producing the overwhelming bias for unidimensional unsupervised classification in the laboratory are the inherent, similarity-based biases that have existed in the stimuli that have been regularly employed. Therefore, I sought to predict when participants should be biased towards unidimensional classification, and when they should be biased towards classification based on a principle of family resemblance, on the basis of the abstract similarity structure of a set of objects. To do this, I employ the simplicity model of unsupervised categorisation (Pothos & Chater, 2002). In Chapter 3 of this thesis, I sought to broaden my investigations of unsupervised categorisation: I examine how stimulus similarity structure influences incidental unsupervised classification in both humans and rats, by either enhancing or limiting the processes of perceptual learning, sensory-preconditioning, and 'surprise'. By investigating incidental unsupervised classification in this way, I sought evidence of human-like categorization behaviour in rats, which some authors would deny (see, e.g., Chater & Heyes, 1994). Finally, Chapter 4 of this thesis investigates whether rats exhibit another important aspect of human categorisation, which some authors have denied (see Chater & Heyes, 1994). That is, Chapter 4 assesses whether rats are capable of engaging in stimulus cross-classification, based on the learned statistical properties of the environment.

Chapter 2

Unidimensional versus two-dimensional classification in human unsupervised categorisation

"...there may be no general answer to the question of which partitioning of some abstract structure of a set of examples is more natural."

(Medin et al., 1987, p. 33)

2. Introduction

When asked to group a set of stimuli in the absence of any feedback, participants readily engage with this task, generating a stimulus classification that, one assumes, is meaningful and intuitive to them. It is apparent, however, that the kind of unsupervised classification behaviour exhibited by participants in the laboratory is at odds with theoretical considerations of the nature of our everyday categories and concepts (see Chapter 1). On the one hand, there exists a general bias among participants to engage in unidimensional unsupervised categorisation in the laboratory (e.g., Ashby et al., 1999; Medin et al., 1987; Regehr & Brooks, 1995), reflecting a reliance on a classical-type approach to categorisation. On the other hand, one sees the theoretical rejection of categorisation based on definitional qualities (i.e., the classical view), due to the fact that our everyday category structures clearly reflect a principle of family resemblance (see Rosch & Mervis, 1975; Wittgenstein, 1953). It is important to ask, therefore, why people choose to sometimes *ignore* some of the dimensions of variation that exist within a stimulus set, and why they often take this to an extreme in laboratory-based studies of unsupervised categorisation.

While unidimensional unsupervised classification dominates in the laboratory, as was shown in Chapter 1, a number of task manipulations have been found to increase family resemblance sorting (e.g., speeded classification; Smith & Kemler Nelson, 1984; cf. Milton & Wills, 2008). Probably the most effective of these manipulations has been incorporating prior knowledge into laboratory-based unsupervised categorisation tasks (e.g., Lassaline & Murphy, 1996; Spalding & Murphy, 1996). The reason for this appears to be that 'knowledge' encourages the interconnection of stimulus features, which enables participants to more readily discover the experimenter-defined family resemblance category structure. While prior knowledge is clearly an important factor in influencing everyday categorisation, models of unsupervised categorisation have typically ignored such factors, defining category coherence purely in terms of similarity (e.g., Pothos & Chater, 2002). Despite ignoring the influence of general knowledge, these models have proved successful in capturing a range of unsupervised categorisation data (see, e.g., Love et al., 2004). Moreover, Love et al. (2004; see also, Ahn & Medin, 1992) have shown that, based on the abstract similarity structure of Medin et al.'s (1987) binary stimulus structure (see Figure 1, Chapter 1), unidimensional classification of the respective stimuli should be considered 'optimal'. While consistent with Medin et al.'s (1987) experimental findings, this prediction is inconsistent with these author's intuitions about this stimulus structure, which they believed would promote classification based on a principle of family resemblance. The fact that this belief has propagated, and that this binary stimulus structure has been so widely employed, has simply compounded the sense that participants are doing something odd, and that they are 'naturally' biased towards unidimensional unsupervised classification. Based on the work of Love et al. (2004), therefore, it is possible that much of the bias for unidimensional unsupervised classification in the laboratory may simply reflect the abstract similarity structure of the stimuli being employed. A number of other factors have also likely contributed to the persistence of unidimensional unsupervised classification in the laboratory (see Section 1.3.2, Chapter 1): these include the specification of the number of categories that should be used for classification, and the almost universal use of binary dimensioned stimuli. This latter point is important because, as Rosch states, "once the S [subject] has learned the rule(s) defining the positive subset, any one stimulus which fits the rule is as good an exemplar of the concept as any other" (1973, p. 329). This does not reflect the nature of real world categories where some stimuli are more typical members of a category than others, and which are often based on stimuli composed of continuous physical variation (Rosch, 1973).

The aim of the present chapter is to investigate the influence of abstract similarity structure on human unsupervised categorisation, in the hope of explaining some of the conflicting results and intuitions presented above, and in Chapter 1. The experimental work presented here is based on the assumption that humans should prefer categories that maximise within-category similarity and minimise between-category similarity, as has been found in the basic level categorisation literature (e.g., Gosselin & Schyns, 2001; Rosch & Mervis, 1975). Furthermore, the focus of the present chapter is on unrestricted unsupervised categorisation, using stimuli composed of continuous physical variation, rather than discrete (binary) variation (although the proposed approach can also, in principle, be applied to stimuli composed of binary dimensions). To be able to appropriately assess the importance of abstract similarity structure in biasing people towards either unidimensional or multidimensional unsupervised classification, it is obviously necessary to be able to establish a means through which one can identify category structure, and assess the intuitiveness of this category structure.

2.1 Assessing category intuitiveness

Given a set of stimuli constructed from *n* dimensions, participants may choose to categorise these stimuli based on just one of the *n* dimensions of variation present, up to a classification based on all *n* dimensions. To make things simple, when considering a set of stimuli constructed from two dimensions of physical variation (*x*,*y*), stimulus classification may proceed in one of three ways: by taking into account dimension *x* only, dimension *y* only, or both dimensions together[4]. Each of these possible dimensionalities will therefore be associated with a different grouping of the stimuli, which I will denote as Group(*x*), Group(*y*) and Group(*x*,*y*), respectively. Critically, these different stimulus groupings will likely differ in their perceived 'naturalness' or 'intuitiveness'. Consider the stimulus structure depicted in Figure 2, for example: when the stimulus points are collapsed along just dimension *x* (Group(*x*)), an obvious ('intuitive') two cluster category structure is formed. In contrast, when the stimulus points are collapsed along just dimension *y* (Group(*y*)), one simply sees homogenous variation along this dimension, and no obvious category structure. When taking into account both dimensions together (Group(*x*,*y*)), a category structure similar to that identified along dimension *x* is apparent, although a lot more variation is introduced by having to consider dimension *y* as well. Consequently, if asked to classify the stimuli depicted in Figure 2, Group(*x*) should be preferred by participants.

[4] For the sake of simplicity, in all the category structures employed in this chapter, dimension *x* and dimension *y* are considered to have equal weighting. It is possible, of course, that two dimensions of variation will not be equally weighted.

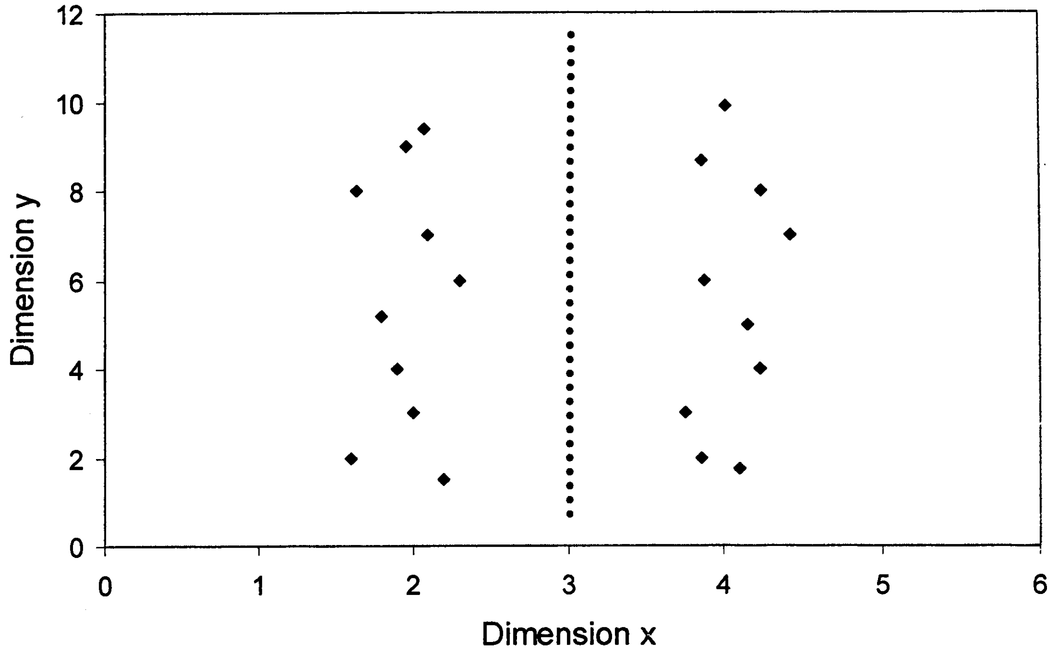

Figure 2. Example stimulus structure where classification along just dimension *x* should be perceived to be 'most intuitive'. Here and elsewhere the dimensions *x* and *y* are assumed to correspond to dimensions of physical variation. In this structure, when the stimuli are represented along just dimension *x*, there is a well-defined two cluster category structure. In contrast, when represented along just dimension *y*, or when taking into account both dimension *x* and dimension *y* together, any category structure is a lot less obvious.

To recapitulate, for stimuli constructed from two dimensions of variation, participants may choose to classify the stimuli by just considering dimension *x* (Group(*x*)), just considering dimension *y* (Group(*y*)), or by considering both dimensions *x* and *y* together (Group(*x*,*y*)). One can ask, therefore, which of these three dimensionalities produces the 'most intuitive' classification (that is, the classification perceived to be 'best' and most obvious)? In making this decision, I first assume that the cognitive system assesses the intuitiveness of Group(*x*) versus Group(*y*) versus Group(*x*,*y*) concurrently (Pomerantz & Kubovy, 1986), and second, that the cognitive system will prefer the dimensionality that produces the most intuitive classification. That is, if the intuitiveness of Group(*x*) (or Group(*y*)) is greater than that of Group(*x*,*y*), then the cognitive system will prefer a unidimensional classification. In contrast, if Group(*x*,*y*) is considered more intuitive, then participants will prefer a two-dimensional categorisation. It is further assumed that these biases will be evident in participants' classification behaviour, and that if the well-formedness of a category structure is not enhanced by the use of additional dimensions, then these dimensions will be ignored in favour of fewer dimensions. So, how can one measure the intuitiveness of Group(x), Group(y), and Group(x,y)?

In Chapter 1, I discussed three influential models of human unsupervised categorisation which are able to identify the 'best' category structure (classification) for a set of stimuli. However, not all of these models are able to provide a comparative assessment of the relative goodness of the three possible classification strategies presented above (i.e., Group(x) versus Group(y) versus Group(x,y)). For example, while the Rational model (Anderson, 1991) may very likely produce different classifications depending on whether stimuli are represented along just dimension x, or through a combination of dimension x and dimension y together, it is not possible to compare the relative goodness of these two classifications. Moreover, the number of categories produced by the Rational model is effectively determined by the model's coupling parameter (that is, the threshold at which new clusters should be formed). As highlighted in Chapter 1, by specifying the number of categories sought, participants may be biased towards employing one classification strategy (e.g., unidimensional classification) over another (Murphy, 2002). Overall, therefore, while alternative Bayesian approaches may be able to capture the issue of unidimensional versus two-dimensional classification (see Cheng, Shettleworth, Huttenlocher, & Rieser, 2007), it is not clear that the Rational model can.

Regarding SUSTAIN (Love et al., 2004), its attentional parameters do allow it to predict a unidimensional versus multidimensional classification preference. Specifically, on the basis of Love et al.'s (2004) simulations, unidimensional unsupervised classification appears to be favoured by SUSTAIN when stimuli are made up of dimensions that do not intercorrelate with each other, or correlate only partially with each other. Order effects in stimulus presentation are critical in determining which dimension is focused upon, and attentional weights are adjusted to favour that clustering (i.e., to make clusters more well-separated). In contrast, two-dimensional classification will be favoured by SUSTAIN when the two dimensions are highly correlated with each other. These predictions of SUSTAIN are supported by the work of Billman and Knutson (1996), who demonstrated the importance of dimensional intercorrelation in unsupervised learning. While SUSTAIN can predict a preference for unidimensional versus multidimensional classification, it provides no quantification (or value) for the intuitiveness of a particular category structure. This quantification is required for the present proposal, where it is necessary to compare Group(x) and Group(x,y). Moreover, the presence of SUSTAIN's cluster recruitment mechanism – that is, the mechanism that specifies the threshold level of dissimilarity required between a novel stimulus and an already formed cluster for that novel stimulus to be accommodated in a new cluster, rather than become part of an existing cluster – indirectly determines the number of categories formed. Similar to the issues surrounding the coupling parameter in the Rational model (Anderson, 1991), therefore, this parameter somewhat confuses the issue of unidimensional versus two-dimensional classification (see Murphy, 2002).

In summary, neither the Rational model nor SUSTAIN appear adequate to assess the influence of abstract similarity structure on participants' preference for unidimensional versus multidimensional classification. In contrast, the simplicity model of unsupervised categorisation (Pothos & Chater, 2002) is ideally suited for this task. To recapitulate, the simplicity model is a computational implementation of Rosch and Mervis's (1975) suggestion that basic level categories maximise within-category similarity and minimise between-category similarity. As outlined in Chapter 1, the simplicity model assesses the gain in 'simplicity' that can be achieved by imposing a specific clustering on a set of stimulus points. The basic premise is that the classification that is deemed the simplest (i.e., 'most intuitive') should be preferred by the participant. Importantly, the simplicity model provides a value for the intuitiveness of a particular classification in terms of an associated *codelength*; shorter (lower value) codelengths are associated with a more intuitive categorisation. Of particular merit is the fact that the simplicity model is *parameter free*, and, when computing category intuitiveness, does not require any specification of the number of categories sought. To investigate the influence of abstract stimulus structure on unidimensional versus multidimensional unsupervised classification, therefore, the simplicity model of unsupervised categorisation (Pothos & Chater, 2002) was employed. Consequently, the experiments reported in this chapter also provide an obvious test of the validity of the simplicity model, although this *was not* the primary goal of these experiments.

2.2 Assessing unidimensional versus two dimensional classification with the simplicity model

To recapitulate, when considering a set of stimuli constructed from two dimensions of physical variation (*x*,*y*), classification may proceed by considering just dimension *x*, just dimension *y*, or both dimensions together (i.e., Group(*x*), Group(*y*), or Group(*x*,*y*), respectively). In determining which of these dimensionalities produces the most intuitive classification for a stimulus set, the simplicity model can be employed to provide an associated codelength value (and clustering pattern) for Group(*x*), Group(*y*), and Group(*x*,*y*). Codelength values are given in terms of a percentage, which represents the number of bits (length of description) required to describe the stimulus sets' similarity information with categories, relative to how many bits are required to describe the same, raw similarity information without categories. The lower a codelength's value, the more intuitive/ natural the classification is considered to be, and the more obvious it should appear to naïve observers (at least, that is the assumption). Consequently, if Codelength(Group(*x*)) or Codelength(Group(*y*)) is less than Codelength(Group(*x*,*y*)), then one would predict that participants should display a *preference for* unidimensional unsupervised classification (henceforth, Codelength(Group(*x*)) is denoted as Codelength(*x*), etc.). Similarly, if Codelength(*x*,*y*) is less than Codelength(*x*) and Codelength(*y*), then one would predict that participants should display a preference for two-dimensional unsupervised classification.[5]

The above paragraph, therefore, specifies one way in which it is possible to assess how abstract similarity structure biases participants' preference for either unidimensional or multidimensional unsupervised classification. It is interesting to note that participants' bias for unidimensional or multidimensional classification has sometimes been considered random (e.g., Medin et al., 1987). However, as highlighted in Chapter 1, a number of factors such as procedural details, stimulus format, and the introduction of prior knowledge (e.g., Lassaline & Murphy, 1996; Milton & Wills, 2004; Milton et al., 2008; Spalding & Murphy, 1996) have all influenced the amount of unidimensional and family resemblance sorting by participants. The experimental work detailed in this chapter, in which classification

[5] Any model that can provide a quantifiable measure of category intuitiveness without information about the number of categories sought would have been equally appropriate to use here.

biases are specified independently of the factors just mentioned, therefore complements the modelling work of Love et al. (2004). To reiterate, based solely on the stimuli's abstract similarity structure, SUSTAIN (Love et al., 2004) predicted that unidimensional classification should be considered 'most intuitive' for the binary stimulus structure of Medin et al. (1987); a prediction that was empirically confirmed. In assessing the validity of the simplicity approach to the problem of unidimensional versus multidimensional classification, therefore, it is interesting to compare this prediction from SUSTAIN with the respective prediction from the simplicity model.

2.3 Examination of some previous findings

In investigations of human unsupervised categorisation, the majority of studies have employed stimulus structures composed from binary dimensions. The most influential stimulus structure of this kind is that originally employed by Medin et al. (1987), depicted in Figure 1 (see Chapter 1). Briefly, Figure 1 shows a four dimensional binary stimulus structure, which species 10 items. When represented along all four binary dimensions, two categories of stimuli are assumed: Category A, which is composed of a category prototype (specified as 1,1,1,1) and four other items that have three features in common with this prototype (e.g., 0,1,1,1), and Category B, which is again composed of a category prototype (specified as 0,0,0,0) and four other items that have three features in common with this prototype (e.g., 1,0,0,0). To reiterate, while Medin et al. (1987; see also Regehr & Brooks, 1995) assumed that this stimulus structure would yield classification based on a principle of family resemblance, across a wide variety of procedures and stimulus formats, they documented a clear bias among participants for classification based on a single dimension (e.g., head shape).

To assess the binary stimulus structure of Medin et al. (1987) with the simplicity model, I assumed that the 1, 0 values reflected coordinates in a multidimensional psychological space. Feature mismatches, which are the main source of similarity information when using binary dimensioned stimuli, can be considered to correspond to the City block distance between vectors; for example, between 0110 and 0100. Accordingly, this method of similarity computation is legitimate for this stimulus structure, and so the City block metric was employed to compute similarities between the different values. When represented along all four dimensions of variation, the predicted optimal classification for the 10 items was that assumed by Medin et al. (1987); that is, the two category structure depicted in Figure 1 (see Chapter 1). The codelength associated with this classification (i.e., Codelength(4d)) was computed to be 94.84% (see Section 1.4.2 of Chapter 1 for an overview of the simplicity model's computational implementation), meaning that this category structure should not be considered particularly obvious and intuitive by participants. By contrast, when specified along just one of the four dimensions of variation, Codelength(1d) was computed to be 51.57%. This codelength indicates that participants should perceive this category structure to be very obvious and intuitive. Consequently, participants should strongly favour classification that takes into account just one of the four stimulus dimensions over classification that takes into account all four stimulus dimensions. As for SUSTAIN (see Love et al., 2004), therefore, when presented with the binary stimulus structure of Medin et al. (1987), the simplicity model readily predicts a preference for unidimensional classification. Moreover, this prediction is based solely on the abstract stimulus structure of the 10 items specified (see Figure 1).

In trying to reduce participants' preference for unidimensional categorisation, Medin et al. (1987; Experiment 4) employed an alternative stimulus set, whereby items were created on the basis of four trinary-valued dimensions (that is, each dimension now had three levels, 0, 1, and 2, representing, for example, a short, medium, and long length of tail). Again, participants were asked to classify the stimuli into two categories. The authors claimed that there existed no straightforward way to divide the items into two groups based on any single stimulus dimension. As for the binary-valued dimensions, in modelling this stimulus set the assumption was made that each trinary dimension corresponds to coordinates in a psychological space, and the City block metric was again employed to compute similarities. This approach induces an ordering in feature values, such that feature 2 is assumed to be 'greater' than feature 1. As the same ordering of feature values is induced in all analyses (unidimensional versus four-dimensional), however, this ordering in feature values should not affect the comparison between Codelength(1d) and Codelength(4d). Ignoring the requirement to classify into two categories (which cannot be modelled within the simplicity approach), Codelength(1d) was computed to be 61.02%, and Codelength(4d) was 56.70%. Therefore, the simplicity model predicts a slight preference for four-dimensional classification in this case. Medin et al. (1987) found that, when presented with this trinary-valued structure, participants were prevented from producing any unidimensional classifications, but equally, they did not produce any classifications based on all four dimensions. While their findings differ slightly from the predictions of the simplicity model, it is important to consider the impact that constraining classification into two categories may have had. To recapitulate, it has been argued that this constraint on classification may encourage unidimensional sorting (Murphy, 2002). Consequently, it seems reasonable to suppose that the upshot of this situation – in which a slight preference for four-dimensional classification is predicted in a situation that should encourage unidimensional classification – will simply be a reduction in the number of unidimensional classifications observed. I would argue, therefore, that the predictions of the simplicity model, which were made on the basis of the abstract stimulus structure employed, are broadly consistent with the findings of Medin et al.'s (1987).

A few unsupervised categorisation studies have also employed stimuli constructed from continuous-valued dimensions, as I will be using in the experiments presented shortly (e.g., Ashby et al., 1999). The simplicity model can similarly be used to assess unidimensional versus multidimensional categorisation in this situation by computing the similarities between points using a Euclidean distance metric. Using a simplified, 20 point version of one of Ashby et al.'s (1999) data sets, containing 10 points along each 'strip' (see Figure 3a), Codelength(x) and Codelength(x,y) were computed. In this case, Codelength(x) was found to be 50.07% and Codelength(x,y) was 80.83%. The simplicity approach, therefore, predicts a clear preference for unidimensional classification along just dimension x. This makes intuitive sense; when all the stimuli are collapsed along dimension x, two extremely well-separated clusters are obvious. In contrast, in the x,y plane, many between-cluster similarities are actually greater than the within-cluster similarities. In line with simplicity's predictions, Ashby et al. (1999) found that participants rapidly came to respond optimally to the two category classification specified along just dimension x in the absence of feedback.

Ashby et al. (1999) also employed a stimulus set in which a two-cluster classification was specified along the diagonal in the x,y plane (see Figure 3b). For this data set, no preference was found either for a two-dimensional or unidimensional classification. Indeed, participants were unable to learn the two-dimensional classification of this structure without feedback. To explain this finding, I created a second data set in which 10 points were specified along each of the two diagonal 'strips' (see Figure 3b). In this case, Codelength(*x*,*y*) was computed to be 81.70%, which is almost identical to Codelength(*x*,*y*) for Figure 3a. This was an expected result: the simplicity model does not take into account the absolute position of points in psychological space, rather it compares pairs of distances by computing whether distance (A, B) is greater than distance (A, C). As such, codelength values are rotationally invariant. In contrast, Codelength(*x*) was computed to be 81.61% (instead of 50.07% for the unrotated Figure 3a), and Codelength(*y*) was 79.53%. Therefore, the simplicity model predicts that, in this case, Group(*x*), Group(*y*) and Group(*x*,*y*) are all, approximately, equally intuitive (although none of the groupings will be particularly obvious), meaning that no one classification should be preferred. This prediction is consistent with the results of Ashby et al. (1999).

As highlighted by the previous example, while rotation does not change the associated two-dimensional codelength values, the unidimensional versus two-dimensional bias can be radically altered. This alteration is caused by the change in 1d projections associated with rotation of the data points. That is, while there is a well-separated unidimensional projection in Figure 3a, this is not the case in Figure 3b. Critically, rotating a data set does not imply that the coordinate axes have to be rotated as well. Rather, the alignment of the coordinate axes is determined by independent, perceptual considerations (a coordinate axis in psychological space can be defined as the direction along which only one aspect of a stimulus' appearance is altered). Consequently, rotation can radically alter the simplicity model's prediction for a unidimensional versus two-dimensional classification bias.

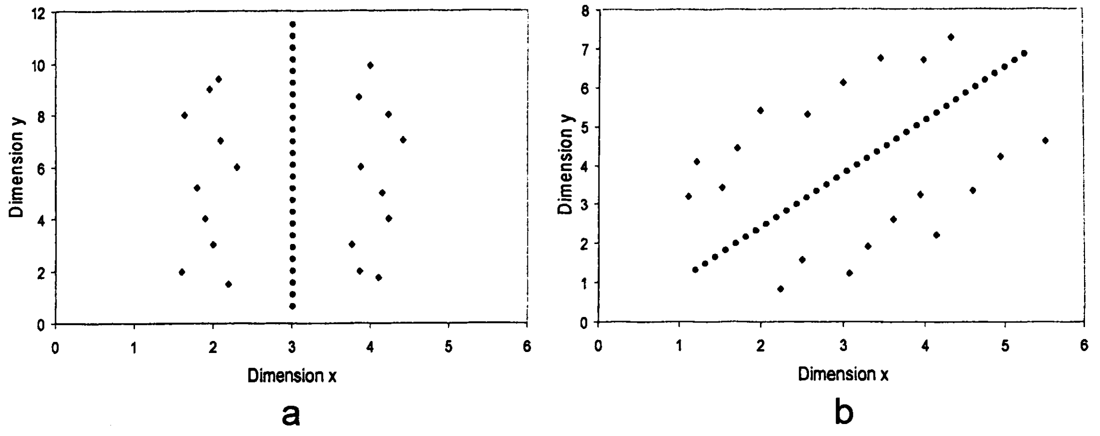

Figure 3. Two simplified versions of the data sets employed by Ashby et al. (1999). For stimulus structure 'a', participants were found to prefer classification along just dimension *x* (the preferred classification is highlighted by the perforated line) rather than classification by taking into account both dimension *x* and dimension *y* together. For stimulus structure 'b', no dimensionality was preferred for classification, and none of Ashby et al.'s (1999) participants responded 'optimally'.

In summary, the predictions derived from the simplicity model are broadly consistent with the findings of Medin et al. (1987; see also Regehr & Brooks, 1995) and Ashby et al. (1999), demonstrating support for the simplicity approach. Critically, these predictions were based solely on the abstract similarity structures of the stimulus sets employed by these authors. With respect to the findings of Medin et al. (1987), the predictions of the simplicity model support those of SUSTAIN (Love et al., 2004), suggesting some level of compatibility between the two models. Before an experimental investigation of the influence of abstract similarity structure on unidimensional versus multidimensional classification can begin, however, two fundamental methodological issues need to be addressed: First, it is important to ensure that the experimental procedure does not bias participants to favour one classification strategy over another (by promoting unidimensional classification, for example). Second, it is clearly necessary to be able to unambiguously infer whether participants are basing their classifications on only one dimension of variation, or on more than one dimension of variation. In Section 2.4, these methodological issues are discussed further, and a procedure is proposed that allows for the assessment of unconstrained unsupervised categorisation.

2.4 Methodological Concerns

A common methodology used in investigations of unsupervised categorisation is to ask participants to classify a set of stimuli into two categories. This has been useful as, in conjunction with the use of binary dimensioned stimuli, it has allowed experimenters to readily assess the basis on which a participants' classification was derived (i.e., on the basis of a single dimension, or multiple dimensions). However, some authors have argued that constraining classification in this manner may inadvertently encourage participants to interpret the experiment as a problem-solving task, rather than a simple categorisation task (Murphy, 2002). Indeed, Murphy (2002) points out that standardised tests in the US often require searching for a critical property to distinguish between instances. Consequently, it is possible that the action of constraining laboratory-based unsupervised categorisation may, in itself, encourage unidimensional classification. Additionally, for a given stimulus set, it is possible that while an intuitive classification into, for example, three categories exists when considering two dimensions of variation, the only intuitive classification that exists into two categories is if participants consider only a single dimension of variation (of course, this could go both ways). So, asking participants to sort a set of stimuli into a particular number of categories may bias their classifications. To adequately examine the influence of abstract similarity structure on the issue of unidimensional versus two-dimensional classification, therefore, an unconstrained categorisation procedure would be preferable. However, there is good reason why previous experimenters have chosen to constrain participants' categorisations.

In unconstrained unsupervised categorisation, there will be considerable response variability: for as few as 10 stimuli, there are about 100,000 possible categorisations (Medin & Ross, 1997). Accordingly, classification performance has to be measured in terms of a person's preference towards one classification, relative to another (e.g., Group(x) versus Group(x,y)). This can be achieved by using a metric of classification similarity, such as the Rand Index (Rand, 1971). The Rand Index is a statistic that can be implemented in categorisation research to compare two classifications. Specifically, it is the number of pairs of stimuli that are both in the same cluster, or both in different clusters, in two classifications, divided by all pairs. It varies from 0 (totally different classifications) to 1 (identical classifications). For example, consider a participant who produces a classification X. Does this classification reflect a unidimensional or a two-dimensional bias? By comparing Rand(X,Group(*x*)) with Rand(X,Group(*x*,*y*)), one can assess this: if, for example, the second Rand is larger, then one can conclude that the participant's classification is more similar to Group(*x*,*y*), indicating that the participant had a bias for two-dimensional classification.

A final issue of concern is the format of the stimuli. In categorisation research, materials are often created in a way that each stimulus can be perceived as an individual object. Sometimes these objects have a naturalistic appearance (e.g., cartoon-like characters, as in Medin et al., 1987), or they correspond to a meaningless geometric shapes (e.g., lines differing in orientation and length, as in Ashby et al., 1999). Regehr and Brooks' (1995) stimuli, for example, were each formed from a separable two-dimensional arrangement of features, such that a stimulus could be composed of a bottle, a cup, a trumpet, and a cake, enclosed within a rectangle. While Milton and Wills (2004; see also, Handel & Imai, 1972) have observed that stimulus format does influence unidimensional versus multidimensional classification, they found it difficult to formulate general principles.

The simplicity approach can only explain biases arising from the abstract stimulus structure of a set of stimuli, not stimulus format or other procedural details. Therefore, the two-dimensional stimuli chosen here were constructed such that they could be perceived as individual objects, as is most commonly the case in categorisation research. However, I also aimed for dimensions of physical variation that would be neither particularly separable nor integral, since this could potentially influence participants' classification preference (Milton & Wills, 2004). Crucially, with the Rand Index analysis, it is not necessary to ensure that the stimulus dimensions do not introduce a bias either for unidimensional or multidimensional classification. Suppose, for example, that the stimulus format encourages a bias for multidimensional classification. Irrespective of this bias, the Rand Index should still reveal *more* of a bias for unidimensional classification in the case where the simplicity model predicts a preference for classification along just dimension *x* (i.e., Group(*x*)), for example, compared to the case where the simplicity model predicts a preference for classification in two dimensions (i.e., Group(*x*,*y*)).

In conclusion, using the simplicity model of unsupervised categorisation (Pothos & Chater, 2002) and the Rand Index (Rand, 1971), one is able to investigate the influence of abstract similarity structure on people's preference for unidimensional versus multidimensional unsupervised classification, in an entirely unconstrained manner. Specifically, one can generate one stimulus structure for which the simplicity model predicts a unidimensional classification bias, and a second stimulus structure for which the simplicity model predicts a two-dimensional classification bias. For both stimulus structures, each dimensionality will be associated with a predicted classification (i.e., Group(x), Group(y), and Group(x,y)). The Rand Index can be used to calculate the similarity of participants' physical classifications of the stimuli to Group(x), Group(y), and Group(x,y). The experiments reported below, therefore, investigate if it is possible to predict unidimensional versus multidimensional classification based on the abstract similarity structure of a set of stimuli.

2.5 Experiment 1

2.5.1 Method

2.5.1.1 Participants

Fifty Cardiff University students took part for course credit. Twenty-five participants were allocated to a condition where a preference for unidimensional classification was predicted, and 25 to a condition where a preference for two-dimensional classification was predicted. A further 24 Cardiff University students participated in a similarity ratings task for course credit.

2.5.1.2 Materials

Stimuli were circles enclosed in squares, with the circles 'blended in' with the squares (using CorelDraw), so as to make them look more like individual objects (see Figure 4). The similarity structure for the two conditions was specified on abstract 1 – 10 scales; as such, these scales had to be applied to the physical dimensions of circle size and square size. This was done by assuming a Weber's fraction of 7.5% for both the circles (smallest size: 24.8 mm) and the squares (smallest size: 52.1 mm; Morgan, 2005). Each stimulus was printed individually on a piece of paper as large as the stimulus, which was subsequently laminated.

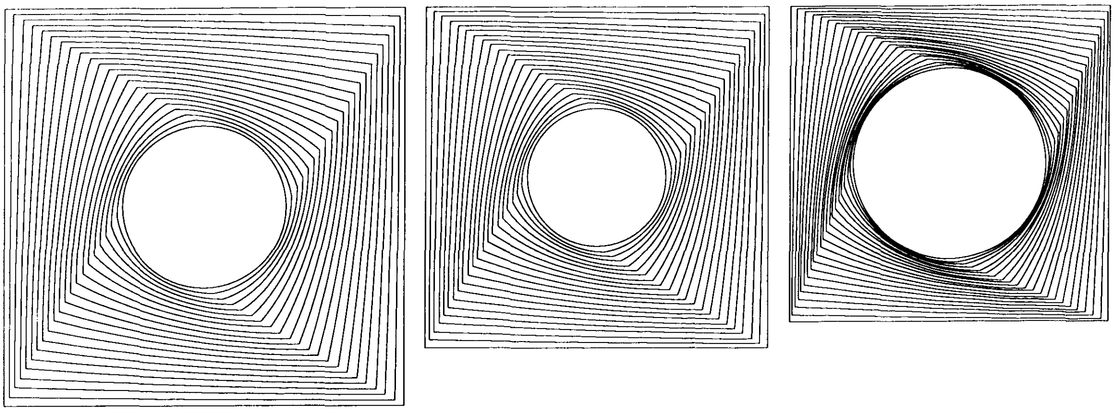

Figure 4. A few examples of the stimuli employed in Experiments 1 to 4. The stimulus presented on the left shows the greatest size in the square dimension, and the stimulus presented on the right shows the greatest size in the circle dimension.

Figures 5 and 6 show the stimulus structures that were used for Experiment 1. Specifically, Figure 5 shows a stimulus structure for which the simplicity model predicts a preference for unidimensional classification, since Codelength(x) and Codelength(y) are less than Codelength(x,y). In two dimensions, there are four, relatively poorly distinguished clusters, whereas along either just x or y, there are two, reasonably well-separated clusters (Group(x) and Group(y) are predicted to be equally intuitive). Figure 6, by contrast, shows a stimulus structure for which the simplicity model predicts a two-dimensional classification preference, since Codelength(x,y) is less than both Codelength(x) and Codelength(y). In two-dimensions, there are two, reasonably well-separated clusters, whereas along either just x or y, there is simply a uniform distribution of stimuli, with no obvious category structure. Importantly, the stimulus sets were created so that the codelengths for the predicted 'optimal' classification(s) in each condition were approximately the same, and likewise for the predicted 'suboptimal' classification(s).

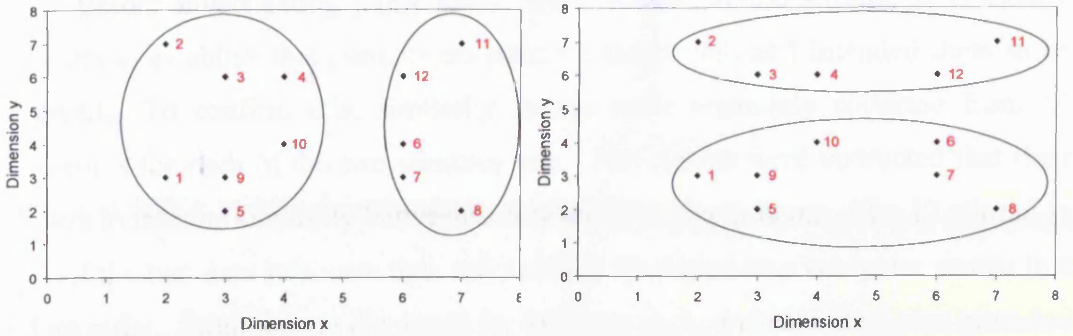

Figure 5. A stimulus structure where the simplicity model predicts a unidimensional classification preference (the unidimensional classifications are shown): the left-hand structure depicts the most intuitive classification along just dimension *x*, in which the predicted 'optimal' clustering is (1,2,3,4,5,9,10) (6,7,8,11,12), and the right-hand structure depicts the most intuitive classification along just dimension *y*, in which the predicted 'optimal' clustering is (1,5,6,7,8,9,10) (2,3,4,11,12). Both these classifications are associated with a codelength of 58.05%. When represented along both dimension *x* and *y* together, the predicted 'optimal' clustering is (1,5,9,10) (2,3,4) (6,7,8) (11,12), with an associated codelength of 76.15%.

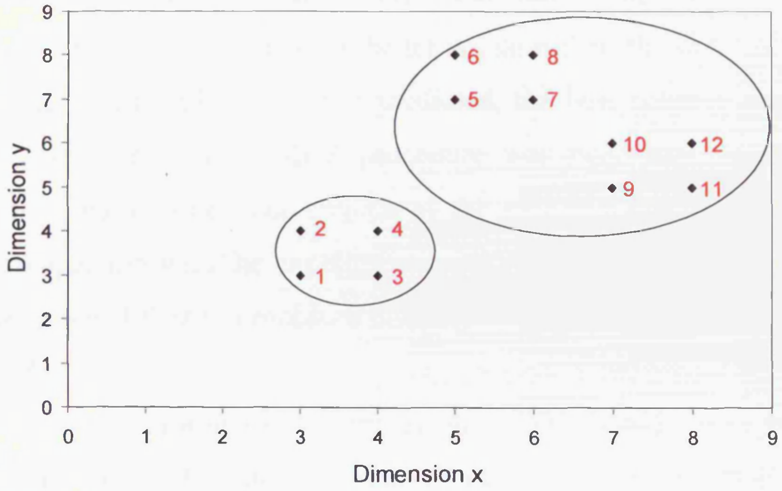

Figure 6. A stimulus structure where the simplicity model predicts a two-dimensional classification preference (the two-dimensional classification is shown): when represented along both dimension *x* and *y* together, the predicted 'optimal' clustering is (1,2,3,4) (5,6,7,8,9,10,11,12), with an associated codelength of 57.40%. Along any single dimension (i.e., either just dimension *x* or just dimension *y*), the predicted 'optimal' clustering is (1,2,3,4) (5,6,7,8) (9,10,11,12), both with an associated codelength of 81.08%.

Before investigating participants' classification of the stimuli, it is clearly important to establish that participants perceive the stimuli as I intended them to be perceived. To confirm this, similarity ratings were separately collected from 12 participants for each of the two stimulus sets. Participants were instructed that their task was to rate the similarity between a number of different items. The 12 stimuli in either of the two data sets were then sequentially displayed on a computer screen in a random order. Stimuli were displayed for 1000 ms each, and each item was preceded by a centrally located fixation point, displayed for 250 ms. Subsequently, participants were instructed that they would have to rate the similarity between the stimuli on a scale ranging from 1 (very dissimilar) to 9 (very similar). Each trial consisted of a central fixation point (250 ms), followed by the first stimulus (1000 ms), followed by another fixation point (250 ms) and the second stimulus (1000 ms), then the similarity scale, which was visible until a response was made. Participants rated the similarity of all possible stimulus pairs once, excluding pairs of identical stimuli, for a total of 132 similarity comparisons. Trials were randomly ordered. Multidimensional Scaling (MDS) was used to derive a spatial representation in two-dimensions for the stimuli, on the basis of participants' similarity ratings. For the data set for which a unidimensional classification bias was predicted, the best solution was associated with a stress of 0.068 (lower values indicate better solution); for the data set for which a two-dimensional classification bias was predicted, the best solution was associated with a stress of 0.097. This MDS procedure was necessary so that a spatial representation of participants' perceptions of the two stimulus structures could be derived for comparison with the experimenter assumed structures. This comparison was made using the Orthosim procedure introduced by Barrett, Petrides, Eysenck and Eysenck (1998).

The Orthosim procedure (Barrett et al., 1998) allows the computation of various similarity indices between two sets of coordinates for the same set of items. By using this procedure, I was able to compare the similarity of the MDS derived representation of the stimuli with the experimenter assumed coordinates (on the basis of which the predictions for unidimensional versus multidimensional classification were computed). A similarity index was used which adopts a 'procrustes' approach (Barrett et al., 1998), according to which the coordinate configurations to be compared are first normalised and rotated/ reflected to remove any of the arbitrariness in MDS solutions (with respect to location, scale, and orientation). The Orthosim documentation recommends the 'double-scaled Euclidean distance' coefficient, for which 0 corresponds to complete dissimilarity, and 1 to identity. The similarity coefficient between the coordinates for the stimulus set where a unidimensional classification bias was predicted and the corresponding MDS solution was 0.79, and for the stimulus set where a two-dimensional classification bias was predicted and the corresponding MDS solution, the similarity coefficient was 0.76. In evaluating the results of the Orthosim procedure, it is important to note that a similarity scale is a rather insensitive measure of similarity perception. Moreover, participants' responding during the ratings task likely became much less careful as the task progressed. Consequently, the similarity ratings procedure can lead to rather noisy data. An alternative procedure for assessing similarity, such as confusability ratings, was not used, as the stimuli employed here are readily discriminable relative to each other. Overall, the similarity between the MDS solutions and the corresponding experimenter assumed coordinates is considered adequate.

2.5.1.3 Procedure

Participants were presented with one of the two stimulus sets and received the following written instructions:

> *"We would like you to simply group the 12 items in a way that feels both natural and intuitive to you. There is no limit to how many groups you can have, but, you should not use more groups than you think is necessary. You may compare the items in any way that you feel will help you, and you are free to change your mind and re-group the items until you are happy."*

Stimuli were presented in a randomly ordered stack, and participants spread the stimuli out on a table to determine their preferred classification by arranging the stimuli into piles.

2.5.2 Results

Of primary interest is participants' preference to engage either in unidimensional or multidimensional classification. Consequently, for each condition, I was interested in assessing the similarity of participants' classifications to Group(x), Group(y), and Group(x,y). As discussed earlier, any analysis that involves frequency of occurrence of different classifications is prohibited due to the large classification variability that will exist (see Medin & Ross, 1997; for completeness, Figure 28 of Appendix 1 shows the frequency with which participants produced classifications based on a specific number of clusters). Therefore, the Rand Index was employed.

As outlined earlier, the Rand Index allows one to infer whether a participant showed a bias for unidimensional classification or multidimensional classification. That is, if a participant preferred unidimensional classification, then the Rand Index (Rand similarity) when comparing that participant's classification to Group(*x*) (or Group(*y*)) will be greater than when comparing that participant's classification to Group(*x*,*y*), and vice versa. Therefore, for all participants in both conditions, I separately computed the Rand similarity between a participant's classification and the respective predictions for Group(*x*), Group(*y*), and Group(*x*,*y*). While it is possible that participants may, for example, prefer Group(*x*) over Group(*y*) if the squares dimension is more salient than the circles one, such differences are not of interest here. Rather, the result of interest concerns whether Group(*x*) or Group(*y*) is preferred over and above Group(*x*,*y*). Consequently, if Rand similarity is greater to Group(*x*) or Group(*y*) than it is to Group(*x*,*y*), then based on this, it is possible to infer a unidimensional classification preference. If Rand similarity is greater to Group(*x*,*y*) than it is to Group(*x*) and Group(*y*), then it is possible to infer a two-dimensional classification preference.

The dependent variable was the similarity of participants' classifications to Group(*x*), Group(*y*), and Group(*x*,*y*), computed using the Rand Index. These computed similarities are denoted as Rand(*x*), Rand(*y*), and Rand(*x*,*y*), respectively, and are shown in Figure 7. A Greenhouse-Geisser corrected Analysis of Variance (ANOVA), with condition (predicted unidimensional preference or predicted two-dimensional preference) as a between-participants factor and Rand similarity (Rand(*x*) or Rand(*y*) or Rand(*x*,*y*)) as a within-participants factor, revealed a significant effect of condition, $F(1, 48) = 9.33$, $p < .005$, no effect of Rand similarity, $F < 1$, and a significant interaction between these factors, $F(1.52, 73.07) = 69.17$, $p < .001$. Tests of simple main effects revealed that there was a significant effect of condition at Rand(*x*), Rand(*y*), and Rand(*x*,*y*) (smallest $F(1, 144) = 24.29$, $p < .001$). Simple main effects further revealed that there was a significant effect of Rand similarity in the condition where a preference for unidimensional classification was predicted (F(1.28, 30.69) = 15.07, p < .001), and in the condition where a preference for two-dimensional classification was predicted (F(1, 24) = 197.88, p < .001).

As can be seen from Figure 7, however, in the condition where a preference for unidimensional classification was predicted, the similarity of participants' classifications to Group(x,y) was significantly greater than to both Group(x) and Group(y) (as assessed with Bonferroni-adjusted paired samples t-tests[6], t(24) = 6.36, p < .001, and, t(19) = 5.70, p < .001, respectively). In the condition where a preference for two-dimensional classification was predicted, the similarity of participants' classifications to Group(x,y) was significantly less than to both Group(x) and Group(y) (as assessed with Bonferroni-adjusted paired samples t-tests, both ts(24) = -14.07, p < .001).

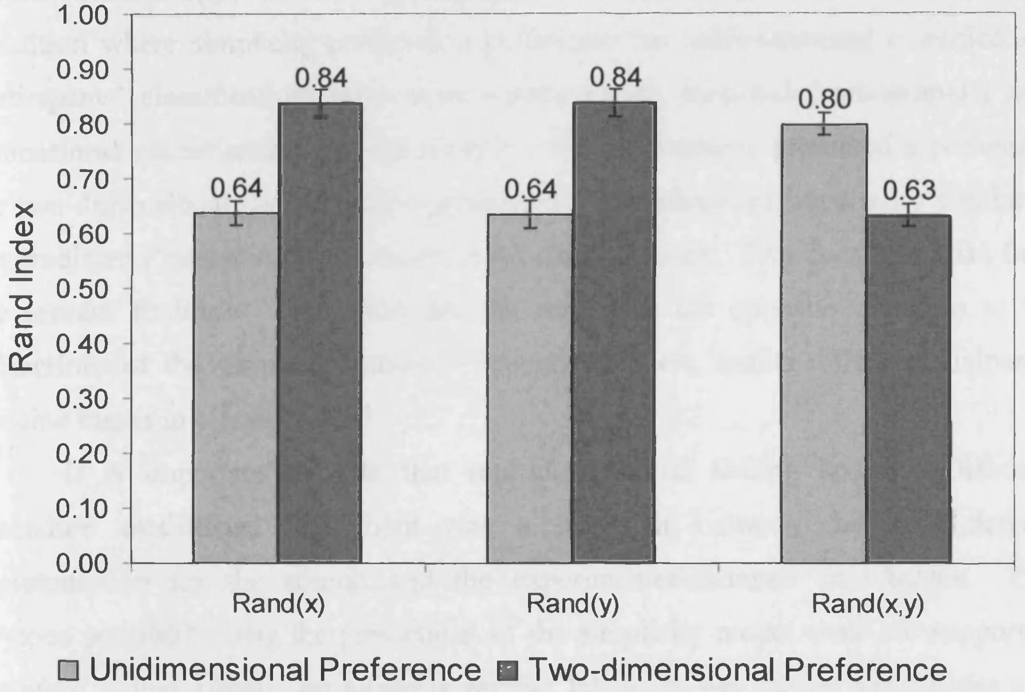

Figure 7. The results of the Rand Index analyses for Experiment 1. Rand(x) means the Rand similarity of participants' classifications to Group(x), etc. 'Unidimensional Preference' refers to the condition where the simplicity model predicted a preference for unidimensional classification. 'Two-dimensional Preference' refers to the condition where the simplicity model predicted a preference for two-dimensional classification. Error bars denote the standard error.

[6] The Bonferroni method of correction has been shown to be extremely robust to violations of sphericity, particularly in terms of controlling Type I error rates (see Field, 2009).

2.5.3 Discussion

Experiment 1 investigated the influence of abstract similarity structure on unidimensional versus multidimensional unsupervised categorisation. Using the simplicity model of unsupervised categorisation (Pothos & Chater, 2002), two stimulus sets were created. For one set of stimuli, a preference for unidimensional classification was predicted, and for the second set of stimuli, a preference for two-dimensional classification was predicted; these predictions were based solely on the abstract similarity structure of the stimuli. The results of Experiment 1 appear to document a preference for unidimensional unsupervised classification, and also the first empirical demonstration of a preference for multidimensional unsupervised classification, on the basis of abstract stimulus structure. However, the pattern of results found is opposite to the predictions of the simplicity model. That is, in the condition where simplicity predicted a preference for unidimensional classification, participants' classifications were more similar to the predicted ('suboptimal') two-dimensional classification. In the condition where simplicity predicted a preference for two-dimensional classification, participants' classifications were more similar to the predicted ('suboptimal') unidimensional classifications. Two questions arise from the present findings: First, why are the results in the opposite direction to the predictions of the simplicity model? Second, do these results reflect participants' genuine biases in classification?

It is important to note that multidimensional scaling and the Orthosim procedure established that there was a good fit between the MDS-derived representation for the stimuli and the experimenter-assumed coordinates. One obvious possibility why the predictions of the simplicity model were not supported, therefore, is that, simply, the model is wrong. However, two further possibilities may also account for participants' classification behaviour being opposite to the predictions of the simplicity model. The first possibility surrounds the idea that, as a result of processing the stimuli, participants may have engaged in some restructuring of similarity space. For example, they may have come to gradually represent the stimuli based on a single, composite, *emergent dimension* along the diagonal.

The second possibility is that the nature of the stimulus structures employed in Experiment 1 may have encouraged what I will term *category subclustering*. That is, for the stimulus sets depicted in Figures 5 and 6, the classifications predicted to be 'suboptimal' (i.e., less intuitive) in each condition held a subordinate relationship with the classifications predicted to be 'optimal' (i.e., more intuitive) in each condition. In each condition, therefore, contained within the 'optimal' (more intuitive) classification structure was meaningful substructure. Consequently, for the stimulus structure depicted in Figure 6, for example, the Rand Index analysis of participants' classification data could have revealed a preference for unidimensional classification in either of two ways: First, participants may have indeed considered unidimensional classification to be 'more intuitive', and so preferred this kind of classification. Second, participants may have considered the two cluster classification in two-dimensions as 'more intuitive' initially, but then sought subclusters in the meaningful substructure along either just dimension x or just dimension y, resulting in an elaboration of the second cluster into two clusters (i.e., (5,6,7,8) and (9,10,11,12)). Indeed, Gosselin and Schyns (2001) have reported that people will often seek to generate classification hierarchies, rather than a single level of classification. Due to the fact that the predicted 'optimal' classification in each condition and the predicted 'suboptimal' classification in each condition shared a superordinate-subordinate relationship, therefore, the Rand Index analysis is unable to determine whether there is a true bias either for unidimensional or two-dimensional classification.

In an attempt to investigate whether the interesting results of Experiment 1 represent true classification biases, an emergent dimension, or category subclustering, Experiments 2, 3 and 4 were undertaken. While Experiment 2 sought to reduce the likelihood of classification based on an emergent dimension, Experiments 3 and 4 sought to reduce the likelihood of any category subclustering.

2.6 Experiment 2

2.6.1 Introduction

Categorisation is obviously dependent on those dimensions that are considered, and when dealing with simple geometric shapes (as employed in Experiment 1), this dependency becomes more acute. While MDS and Orthosim reported an adequate fit between the experimenter-assumed coordinates and the MDS-derived representations for the stimulus structures, as highlighted in the previous section, it is possible that participants' classifications in Experiment 1 may have been influenced by an unanticipated emergent dimension. In an attempt to counter this possibility, Experiment 2 sought to focus participants' attention on the stimuli's 'relevant' dimensions of variation (i.e., the size of the inner circle and the size of the square). By increasing the saliency of the 'relevant' dimensions, this should reduce the likelihood of classification being determined by some unanticipated emergent dimension. Consequently, if participants in Experiment 2 show the same pattern of classification behaviour as participants in Experiment 1, then one can be more confident that the results of Experiment 1 were reflective of classification determined by the 'relevant' dimensions of variation, and *not* some unanticipated emergent dimension. Moreover, to get a better sense about which dimension(s) of variation participants were basing their classifications on, at the end of classification, participants were asked to describe how and why they grouped the stimuli in the way that they did.

2.6.2 Method

2.6.2.1 Participants, materials and procedure

Forty Cardiff University students took part for a payment of £2. Twenty participants were allocated to a condition where a preference for unidimensional classification was predicted, and 20 to a condition where a preference for two-dimensional classification was predicted. The same materials and procedure used in Experiment 1 were employed with the following exception: the instructions presented to participants before the classification task now highlighted the 'relevant' dimensions of variation. Specifically, the instructions read as follows:

> *"We would like you to simply group the 12 items in a way that feels both natural and intuitive to you. There is no limit to how many groups you can have, but, you should not use more groups than you think is necessary. You may compare the items in any way that you feel will help you, and you are free to change your mind and re-group the items until you are happy.*
>
> *Following grouping, you will be asked to describe how and why you grouped the stimuli in the way that you did. For example, you may have based your grouping just on the overall size of the stimulus squares, or just on the size of the inner circles. Alternatively, you may have based your grouping on a*

combination of both the overall size of the stimulus squares and the size of the inner circles."

2.6.3 Results

As for Experiment 1, the dependent variable was the similarity of participants' classifications to Group(x), Group(y), and Group(x,y), as computed using the Rand Index. These computed similarities are denoted as Rand(x), Rand(y), and Rand(x,y), respectively, and are shown in Figure 8 (Figure 29 of Appendix 1 shows the frequencies with which participants produced classifications based on a specific number of clusters). Greenhouse-Geisser corrected ANOVA, with condition (predicted unidimensional preference or predicted two-dimensional preference) as a between-participants factors and Rand similarity (Rand(x) or Rand(y) or Rand(x,y)) as a within-participants factor, revealed a significant effect of condition, $F(1, 38) = 22.61$, $p < .001$, no effect of Rand similarity, $F(1.35, 51.13) = 2.92$, $p > .05$, and a significant interaction between these factors, $F(1.35, 51.13) = 34.20$, $p < .001$. Tests of simple main effects revealed that there was a significant effect of condition at Rand(x), Rand(y), and Rand(x,y) (smallest $F(1, 114) = 8.29$, $p < .005$). Simple main effects further revealed that there was a significant effect of Rand similarity in the condition where a preference for unidimensional classification was predicted ($F(1.16, 22) = 6.87$, $p < .015$), and in the condition where a preference for two-dimensional classification was predicted ($F(1, 19) = 146.14$, $p < .001$).

The results of Experiment 2 are in line with the findings of Experiment 1: that is, in the condition where a preference for unidimensional classification was predicted, the similarity of participants' classifications to Group(x,y) was significantly greater than to both Group(x) and Group(y) (as assessed with Bonferroni-adjusted paired samples t-tests, $t(19) = 2.73$, $p < .015$, and, $t(19) = 4.77$, $p < .001$, respectively). In the condition where a preference for two-dimensional classification was predicted, the similarity of participants' classifications to Group(x,y) was significantly less than to both Group(x) and Group(y) (as assessed with Bonferroni-adjusted paired samples t-tests, both ts$(19) = -12.09$, $p < .001$).

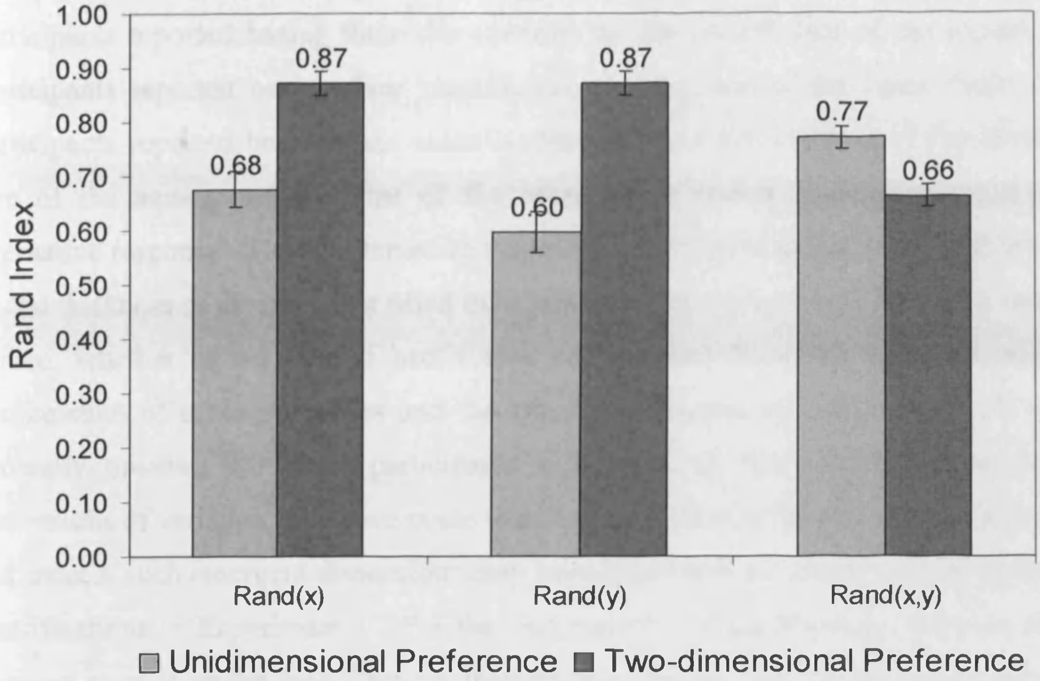

Figure 8. The results of the Rand Index analyses for Experiment 2. Rand(*x*) means the Rand similarity of participants' classifications to Group(*x*), etc. 'Unidimensional Preference' refers to the condition where the simplicity model predicted a preference for unidimensional classification. 'Two-dimensional Preference' refers to the condition where the simplicity model predicted a preference for two-dimensional classification. Error bars denote the standard error.

2.6.3 Discussion

Experiment 2 replicates the findings of Experiment 1: in the condition where simplicity predicted a preference for unidimensional classification, participants' classifications were most similar to the predicted ('suboptimal') two-dimensional classification. In the condition where simplicity predicted a preference for two-dimensional classification, participants' classifications were most similar to the predicted ('suboptimal') unidimensional classifications. Based on the reasoning outlined earlier, I take this replication to provide evidence in support of the view that participants' classifications were unlikely to be (primarily) based on some unanticipated emergent dimension, rather than the experimenter-assumed dimensions. In support of this, an assessment of participants' descriptions about how they decided to classify the stimuli was broadly consistent with the view that participants were focusing on the experimenter-assumed dimensions of variation. Overall, 15 participants reported basing their classification on the overall size of the square, 9 participants reported basing their classification on the size of the inner circle, 11 participants reported basing their classification on some combination of the overall size of the square and the size of the inner circle, and 5 participants gave an alternative response. These 'alternative responses' included basing their classification on the thickness of the lines that filled the space between the inner circle and the outer square, whether or not stimuli had a dark edge around the inner circle, or some combination of these properties and the experimenter-assumed dimensions. While generally positive, for some participants it is apparent that other, unanticipated dimensions of variation may have come to influence their classification of the stimuli; and indeed, such emergent dimensions may have influenced a number of participants' classifications in Experiment 1. For the vast majority of participants, however, this does not seem to be the case. Instead, they reported basing their classifications on the experimenter-assumed dimensions. As a caveat, this qualitative analysis of participants' classification behaviour must be treated with caution, as it is quite possible that the manner in which participants thought they had classified the stimuli did not actually reflect the true manner in which they did classify the stimuli (hence why the Rand Index was employed in the first place).

Given the results of Experiment 2, Experiments 3 and 4 sought to reduce the possibility that participants would engage in category subclustering when presented with the stimuli depicted in either Figure 5 or Figure 6. To recapitulate, due to the superordinate-subordinate relationship that exists between the predicted 'optimal' classification(s) and the predicted 'suboptimal' classification(s) in each stimulus structure, it is possible that while participants may have initially engaged in classification in the manner predicted by the simplicity model, subsequently they may have sought subclusters in the meaningful substructure. A number of factors may have contributed to this possible category subclustering behaviour. First, there is the small number of stimuli to be classified in each condition. Experimenter observations found that classification of the stimuli was rather quick (e.g., between one and two minutes). Given that many psychological experiments that undergraduate participants participate in are rather lengthy (e.g., between 15 – 30 minutes), the shortness of the classification task may have encouraged them to seek subclusters in an effort to demonstrate that they had fully engaged in the classification task. In an attempt to reduce the likelihood of this, participants in Experiment 3 were asked to classify double the number of stimuli than participants in Experiments 1 and 2. A second factor that may have contributed to possible category subclustering is the unlimited amount of time that participants had in which to complete their classifications. Again, this may have encouraged participants to seek category subclusters to demonstrate that they had fully engaged in the classification task. Principally, category subclustering will necessarily take longer to engage in than classification based on one's initial preference. Consequently, in Experiment 4 I introduced a strict time constraint on classification in attempt to reduce any possibility of category subclustering.

2.7 Experiment 3

2.7.1 Introduction

Experiment 3 assessed whether doubling the number of stimuli to be classified would encourage participants to classify the stimuli of Figures 5 or 6 in a manner that is consistent with the predictions of the simplicity model, by reducing any tendency to engage in category subclustering. Why would this manipulation reduce category subclustering? As noted earlier, the small number of stimuli used in Experiments 1 and 2 meant that participants completed their classification of the stimuli rather quickly. Consequently, participants may have tried to do more with their classifications than they would have done 'naturally' (i.e., engaged in subclustering), so as to engender the sense that they had performed adequately in the task. By doubling the number of stimuli to be classified, therefore, this should make the task more effortful and also increase the amount of time that it takes for participants to complete their classifications. As a result of this, participants should feel less pressure to do more with their classifications than they would have done 'naturally'. Moreover, by increasing the number of stimulus comparisons that must be made, it is possible that this may enhance the perceived structural regularities contained within each stimulus set. In doing so, this may help to promote a sense that one stimulus classification – that is, the classification predicted to be 'more intuitive' by the simplicity model – is 'optimal' relative to the other classification possibilities. Of course, it is also possible that this may enhance any category subclustering.

2.7.2 Method

2.7.2.1 Participants

Forty Cardiff University students took part for a small payment of £2. Twenty participants were allocated to a condition where a preference for unidimensional classification was predicted, and 20 to a condition where a preference for two-dimensional classification was predicted.

2.7.2.2 Materials and procedure

The same instructions and materials employed in Experiment 1 were used in Experiment 3. However, for every data point (see Figures 5 and 6), two identical stimuli were generated, creating a total of 24 stimuli to be classified in each condition. While the predicted clustering patterns for each stimulus set do not change from those detailed in Experiment 1 (for example, simplicity still predicts classification into two clusters for Group(x) and Group(y) in Figure 5), due to the increased number of pairwise similarity comparisons, Codelength(x), Codelength(y), and Codelength(x,y) in each condition are altered slightly. Critically, however, doubling the number of stimuli to be classified does not affect the overall unidimensional versus multidimensional classification predictions. In the case where a preference for unidimensional classification is predicted, Codelength(x) and Codelength(y) are now 56.15%, and Codelength(x,y) is now 75.1%. In the case where a preference for two-dimensional classification is predicted, Codelength(x) and Codelength(y) are now 78.6%, and Codelength(x,y) is now 54.7%.

2.7.3 Results

The results of interest are presented in Figure 9 (see Figure 30 of Appendix 1 for the frequency with which participants produced classifications based on a specific number of clusters). Again, the dependent variable was the similarity of participants' classifications to Group(x), Group(y), and Group(x,y), as computed using the Rand Index (denoted as Rand(x), Rand(y), and Rand(x,y), respectively). Inspection of Figure 9 reveals that the pattern of results for Experiment 3 is very similar to that of Experiment 1 and Experiment 2. That is, in the case where the simplicity model predicted a preference for unidimensional classification, participants' classifications are most similar to the predicted 'suboptimal' two-dimensional classification. In the case where simplicity predicted a preference for two-dimensional classification, participants' classifications are most similar to the predicted 'suboptimal' unidimensional classifications.

Greenhouse-Geisser corrected ANOVA, with condition (predicted unidimensional preference or predicted two-dimensional preference) as a between-participants factors and Rand similarity (Rand(*x*) or Rand(*y*) or Rand(*x*,*y*)) as a within-participants factor, revealed a significant effect of condition, $F(1, 38) = 8.86, p < .006$, an effect of Rand similarity, $F(1.47, 55.66) = 4.59$, $p < .015$, and a significant interaction between these factors, $F(1.47, 55.66) = 55.93$, $p < .001$. Tests of simple main effects revealed that there was a significant effect of condition at Rand(*x*), Rand(*y*), and Rand(*x*,*y*) (smallest $F(1, 114) = 15.18$, $p < .001$). Simple main effects further revealed that there was a significant effect of Rand similarity in the condition where a preference for unidimensional classification was predicted ($F(1.34, 25.44) = 15.43$, $p < .001$), and in the condition where a preference for two-dimensional classification was predicted ($F(1, 19) = 260.37, p < .001$).

Consistent with the results of Experiments 1 and 2, in the condition where a preference for unidimensional classification was predicted, the similarity of participants' classifications to Group(*x*,*y*) was significantly greater than to both Group(*x*) and Group(*y*) (as assessed with Bonferroni-adjusted paired samples t-tests, $t(19) = 2.76$, $p < .015$, and, $t(19) = 8.29$, $p < .001$, respectively). In the condition where a preference for two-dimensional classification was predicted, the similarity of participants' classifications to Group(*x*,*y*) was significant less than to both Group(*x*) and Group(*y*) (as assessed with Bonferroni-adjusted paired samples t-tests, both *t*s(19) $= -16.14, p < .001$).

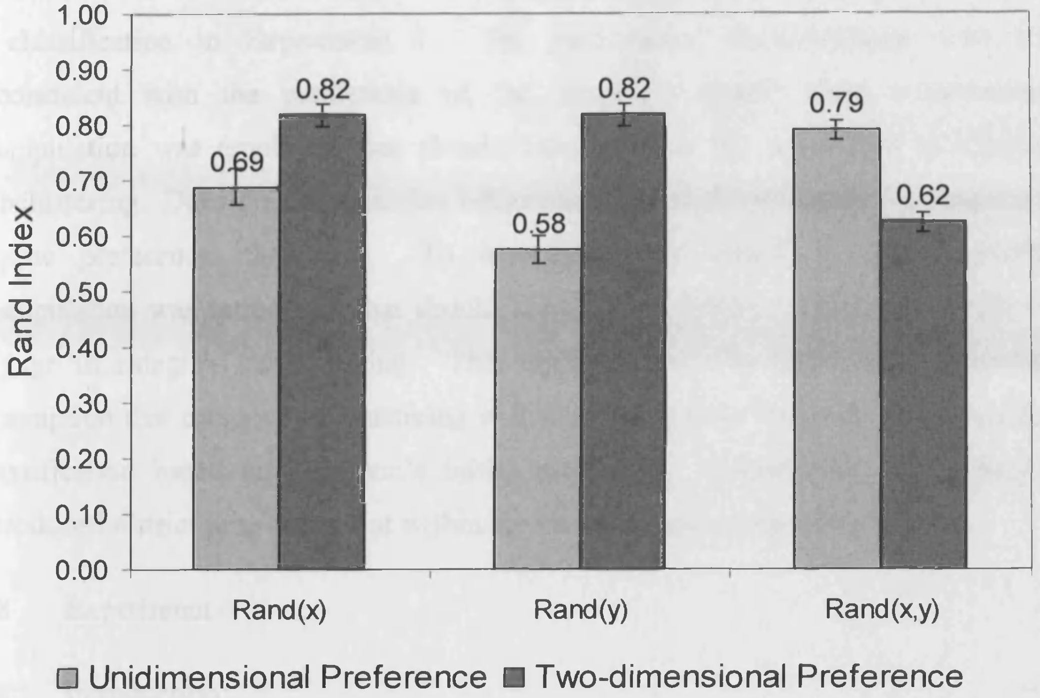

Figure 9. The results of the Rand Index analyses for Experiment 3. Rand(*x*) means the Rand similarity of participants' classifications to Group(*x*), etc. 'Unidimensional Preference' refers to the condition where the simplicity model predicted a preference for unidimensional classification. 'Two-dimensional Preference' refers to the condition where the simplicity model predicted a preference for two-dimensional classification. Error bars denote the standard error.

2.7.4 Discussion

The pattern of results in Experiment 3 was identical to that found in Experiments 1 and 2. That is, in the condition where simplicity predicted a preference for unidimensional classification, participants' classifications were most similar to the predicted ('suboptimal') two-dimensional classification. In the condition where simplicity predicted a preference for two-dimensional classification, participants' classifications were most similar to the predicted ('suboptimal') unidimensional classification.

Overall, therefore, doubling the number of stimuli to be classified in each condition did little to encourage classification that was consistent with the predictions of the simplicity model. However, it is noteworthy that in the condition where simplicity predicted a unidimensional classification preference, two participants produced a classification that matched exactly Group(*x*); no participant produced such a classification in Experiment 1. So, participants' classifications were still inconsistent with the predictions of the simplicity model when a procedural manipulation was employed that should have reduced the possibility of category subclustering. Does the classification behaviour found in Experiments 1 – 3 represent a true preference, therefore? To investigate this further, a more powerful manipulation was introduced that should considerably hinder participants' ability to engage in category subclustering. This manipulation was based on the intuitive assumption that category subclustering will necessarily take longer to engage in than classification based on a person's initial preference. Specifically, Experiment 4 introduced a strict time constraint within the unsupervised categorisation task.

2.8 Experiment 4

2.8.1 Introduction

As noted in Chapter 1, speeded categorisation has been found to both increase (Smith & Kemler Nelson, 1984) and decrease (Milton, et al., 2008) family resemblance sorting. Interestingly, the category structure employed by Milton et al. (2008) was that of Medin et al.'s (1987), for which the simplicity model predicted a preference for unidimensional classification. It seems plausible to suppose, therefore, that when under pressure to categorise a set of stimuli in a short period of time, participants will resort to classification that represents the 'most intuitive' classification based on the abstract similarity structure of the stimuli (this certainly makes intuitive sense). With respect to the results of Experiments 1 to 3, therefore, introducing a tight time-constraint on participants' classifications should greatly encourage them to classify the stimuli on the basis of their initial preference (because they will have little time to engage in category subclustering). Consequently, the similarity of participants' classifications to the predicted classifications of the simplicity model (as computed by the Rand Index) should reflect a more accurate assessment of participants' true classification biases. That is, by forcing participants to classify the stimuli of Experiment 1 rapidly, it was assumed that this would only allow classification at the 'more intuitive' level (i.e., at the assumed 'basic level' of classification). In Experiment 4, therefore, I would argue that the Rand Index can be used to infer participants' classification preferences, even when the predicted 'optimal' classification shares a superordinate-subordinate relationship with the predicted 'suboptimal' classification.

2.8.2 Method

2.8.2.1 Participants and materials

Thirty Cardiff University students took part for a small payment of £2. Fifteen participants were allocated to a condition where a preference for unidimensional classification was predicted, and 15 to a condition where a preference for two-dimensional classification was predicted.

2.8.2.2 Materials

While the same materials of Experiment 1 were used in Experiment 4, the instructions differed from those of Experiment 1. In the present experiment I wanted classification to be as rapid as possible. Based on the findings of an informal pilot study ($N = 5$), I concluded that a classification time of 10 seconds represented a good trade off between classification that was very rapid, but still achievable. Consequently, participants read the following instructions:

> *"We would like you to simply group the 12 items in a way that feels both natural and intuitive to you. There is no limit to how many groups you can have, but, you should not use more groups than you think is necessary. You may compare the items in any way that you feel will help you, and you are free to change your mind and re-group the items until you are happy.*
>
> *You will, however, only have 10 seconds to complete your grouping of the stimuli."*

2.8.2.3 Procedure

After reading the instructions, the experimenter reiterated to the participant that they would have just 10 seconds in which to classify the 12 stimuli. Participants were further told that the experimenter would tell them when to start, and that they would count down the final five seconds of the task (i.e., 5, 4, 3, 2, 1, STOP).

2.8.3 Results

The results from Experiment 4 are presented in Figure 10 (see Figure 31 of Appendix 1 for the frequency with which participants produced classifications based on a specific number of clusters). Importantly, all participants successfully completed the classification task within the allotted time (although this was a struggle for many). The present manipulation influenced participants' classification behaviour in a number of interesting ways: First, in the case where simplicity predicted a preference for unidimensional classification, the similarity of participants' classifications to Group(x) and Group(x,y) was now equivalent (Rand(x) = 0.70, and, Rand(x,y) = 0.72). This result sits in notable contrast to the findings of Experiments 1 – 3. Second, relative to the results of Experiments 1 – 3, in the case where simplicity predicted a preference for two-dimensional classification, there was a marked reduction in the difference between the similarity of participants' classification to Group(x)/ Group(y) and Group(x,y); however, the results still showed that, overall, participants' classifications were still more similar to the predicted 'suboptimal' unidimensional classification.

An ANOVA, with condition (predicted unidimensional preference or predicted two-dimensional preference) as a between-participants factors and Rand similarity (Rand(x) or Rand(y) or Rand(x,y)) as a within-participants factor, revealed no effect of condition, $F(1, 28) = 2.81, p > .05$, a significant effect of Rand similarity, $F(2, 56) = 4.38, p < .02$, and a significant interaction between these factors, $F(2, 56) = 14.47, p < .001$. Tests of simple main effects revealed that there was a significant effect of condition at Rand(y) ($F(1, 84) = 19.12, p < .001$), but not at Rand($x$) or Rand($x,y$) ($F(1, 84) = 1.12, p > .05$, and, $F(1, 84) = 2.77, p > .05$, respectively). Simple main effects further revealed that there was a significant effect of Rand similarity in the condition where a preference for unidimensional classification was predicted ($F(2, 56) = 12.91, p < .001$), and in the condition where a preference for two-dimensional classification was predicted ($F(2, 56) = 5.46, p < .007$).

Focusing on the condition where simplicity predicted a unidimensional classification preference, follow-up tests revealed that while the similarity of participants' classifications to Group(x,y) was significantly greater than to Group(y) (as assessed with a Bonferroni-adjusted paired samples t-test, $t(14) = 4.91, p < .001$), such a difference was *not* found when comparing the similarity of participants' classifications to Group(x,y) and Group(x) (assessed in the same way, $t(14) = .48$, $p > .05$). The latter result sits in contrast to the findings of Experiments 1 – 3, where the similarity of participants' classifications to Group(x,y) was also greater than to Group(x). The fact that participants' classifications were now equally similar to Group(x,y) and Group(x) in this condition is, therefore, rather interesting. Indeed, one may argue that it lends some validity to the argument that part of the reason why people's classification preferences do not match those predicted by the simplicity model in Experiments 1 – 3 is because of a tendency to engage in category subclustering. However, it must be noted that in the condition where simplicity predicted a preference for unidimensional classification, the present manipulation still did not produce a result that was consistent with the predictions of the simplicity model. Moreover, in the condition where simplicity predicted a preference for two-dimensional classification, the similarity of participants' classifications to Group(x,y) was still significantly less than to both Group(x) and Group(y) (as assessed with Bonferroni-adjusted paired samples t-tests ($t(14) = -3.43$, $p < .005$, and $t(14) = -3.43$, $p < .005$, respectively). While this latter result agrees with the findings of Experiments 1 – 3, it is interesting to note that the estimated effect sizes for these differences in Experiment 4 (rs = -0.68) is significantly reduced compared to the same estimated effect sizes in Experiment 1 (rs = -0.94, $Z = 2.53$, $p < .015$; Rosenthal, 1991).

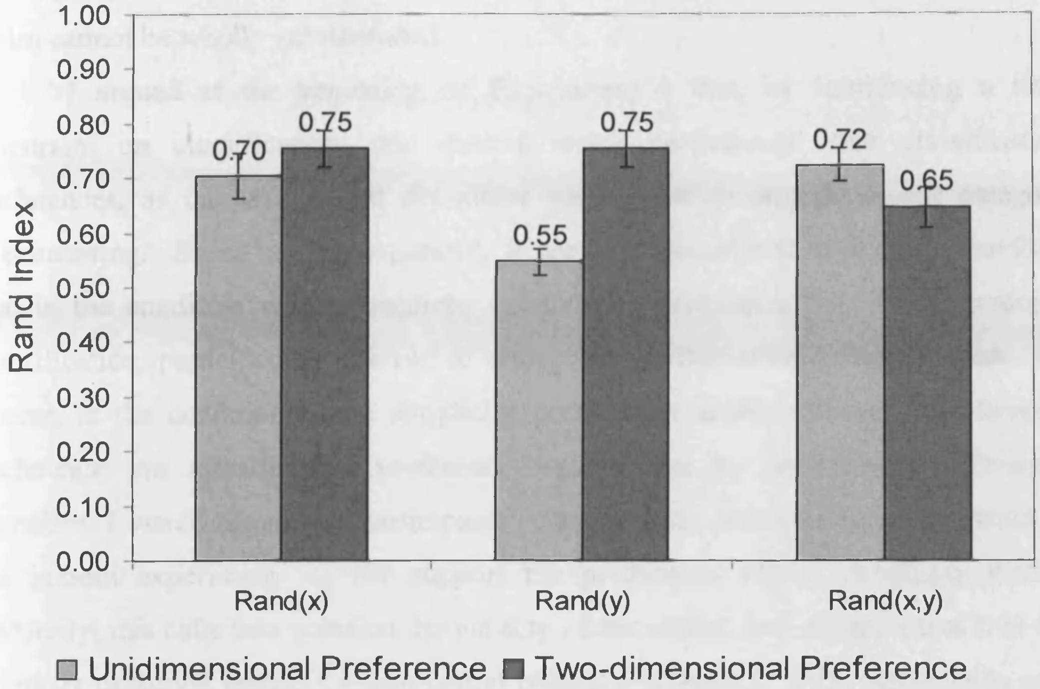

Figure 10. The results of the Rand Index analyses for Experiment 4. Rand(*x*) means the Rand similarity of participants' classifications to Group(*x*), etc. 'Unidimensional Preference' refers to the condition where the simplicity model predicted a preference for unidimensional classification. 'Two-dimensional Preference' refers to the condition where the simplicity model predicted a preference for two-dimensional classification. Error bars denote the standard error.

2.8.4 Discussion

Broadly, the pattern of results of Experiment 4 is consistent with the pattern of results of Experiments 1 – 3 (i.e., in the sense that they are not consistent with the predictions of the simplicity model). However, in the condition where a preference for unidimensional classification was predicted, the similarity of participants' classifications to Group(*x*) was found to be equivalent to that of Group(*x*,*y*). As mentioned above, this finding sits in contrast to the results of Experiments 1 – 3. In the condition where a preference for two-dimensional classification was predicted, participants' classifications were still more similar to the predicted 'suboptimal' (unidimensional) classifications. While the former results (i.e., in the predicted unidimensional classification condition) may lend some validity to the argument that participants in Experiments 1 – 3 engaged in category subclustering, overall, this claim cannot be wholly substantiated.

I argued at the beginning of Experiment 4 that, by introducing a time constraint on classification, this should reveal participants' true classification preferences, as the task would not allow participants to engage in any category subclustering. Based on this argument, it seems reasonable to conclude, therefore, that in the condition where simplicity predicted a preference for two-dimensional classification, participants *preferred* to engage in unidimensional classification. Of course, in the condition where simplicity predicted a unidimensional classification preference, no classification preference was shown by participants. Overall, therefore, I would argue that participants' classification preferences, documented in the present experiment, *do not* support the predictions of the simplicity model. Naturally, this calls into question the validity of the model, and suggests that it is not correctly capturing people's classification biases/ preferences. It is important to note that some of the differences observed between Experiments 1 – 3 and Experiment 4 are likely the result of the classification data simply being much noisier in Experiment 4, due to the time constraint imposed. However, there is no reason to believe that this noise should have acted directly against the predictions of the simplicity model.

Irrespective of what the results mean for the validity of the simplicity model, the results of Experiment 4 *do* document a further interesting influence of time pressure on human unsupervised classification (see, e.g., Milton et al., 2008). Indeed, taken as a whole, the pattern of results of Experiment 4 are particularly interesting because they show that speeded categorisation does not, necessarily, influence people's classifications in a uniform manner (i.e., by simply increasing unidimensional classification, for example; Milton et al., 2008). Rather, speeded categorisation seems to influence people's classification strategies in an apparently more complex manner than has been previously assumed (with respect to the present stimulus structures, at least; cf. Milton et al., 2008; Ward, 1983). That is, relative to the findings of Experiment 1, in the case where a preference for unidimensional classification was predicted, speeded classification increased the similarity of participants' classifications to Group(x) (Rand(x) = 0.64, Experiment 1; Rand(x) = 0.70, Experiment 4), and decreased the similarity of participants' classifications to Group(x,y) (Rand(x,y) = 0.80, Experiment 1; Rand(x,y) = 0.72, Experiment 4). In contrast, in the case where a preference for two-dimensional classification was predicted, speeded classification slightly increased the similarity of participants' classifications to Group(*x*,*y*) relative to Experiment 1 (Rand(*x*,*y*) = 0.63, Experiment 1; Rand(*x*,*y*) = 0.65, Experiment 4), and it decreased the similarity of participants' classifications to Group(*x*)/ Group(*y*) relative to Experiment 1 (Rand(*x*)/ Rand(*y*) = 0.84, Experiment 1; Rand(*x*)/ Rand(*y*) = 0.75, Experiment 4). In general, however, it does appear as though unidimensional classification is somewhat more robust than classification based on a family resemblance principle (cf. Milton et al., 2008).

In conclusion, the speeded classification task of Experiment 4 did not encourage participants to show classification behaviour that was consistent with the predictions of the simplicity model. This is problematic for the simplicity model, as given the short period of time in which participants had to classify the stimuli, it seems reasonable to assume that participants' classification strategies were reflective of a true preference. That is, the speeded classification task of Experiment 4 makes the possibility of category subclustering much less likely, although, of course, still possible. While an even stronger task manipulation (for example, combining speeded classification with a task that places a high demand on working memory) may lead to a reversal in the pattern of results of Experiments 1 – 4, the simplicity model is supposed to capture human classification in the absence of such manipulations. The fact is, although category subclustering is compatible with the simplicity model, the model clearly predicts that such classification is 'suboptimal' given the category structures of Figures 5 and 6. Whether the results of Experiments 1 – 4 indicate participants' true biases or not, these experiments have shown that participants' final classifications have been consistently more similar to the predicted 'suboptimal' classifications than to the predicted 'optimal' classifications. To put the results of Experiments 1 – 4 into context, it is clearly important to establish whether the simplicity model adequately captures participants' classification preferences when stimulus structures are employed in which the predicted 'optimal' classifications do not share a superordinate-subordinate relationship with the predicted 'suboptimal' classifications. Consequently, this was the focus of the experimental investigation in Experiment 5.

2.9 Experiment 5[7]

2.9.1 Introduction

Experiment 5 sought to assess participants' unsupervised categorisation behaviour using stimuli derived from stimulus structures in which the predicted 'optimal' and 'suboptimal' classifications do not share a superordinate-subordinate relationship. Consequently, two new stimulus structures were generated (see Figures 12 and 13). As for Experiment 1, for one of these stimulus structures, the simplicity model predicted a unidimensional classification preference; for the other stimulus structure, the simplicity model predicted a two-dimensional classification preference. Critically, in the case where a unidimensional classification preference was predicted, care was taken to ensure that Group(x,y) was not subordinate to Group(x)/ Group(y), and in the case where a two-dimensional classification preference was predicted, care was taken to ensure that Group(x)/ Group(y) were not subordinate to Group(x,y). Essentially, for both stimulus structures, the category structure that corresponds to classification along a single dimension of variation was made as different as possible to the category structure that corresponds to classification when taking into account both dimensions of variation together. If category subclustering produced the conflict between the predictions of the simplicity model and the experimentally observed classification behaviour of participants in Experiments 1 – 4, then Experiment 5 should elicit classification behaviour that is consistent with the predictions of the model.

2.9.2 Method

2.9.2.1 Participants

Forty Cardiff University students took part for course credit. Twenty participants were allocated to a condition where a preference for unidimensional classification was predicted, and 20 to a condition where a preference for two-dimensional classification was predicted. As for Experiment 1, an additional 24 Cardiff University students provided similarity ratings for a small payment of £2.

[7] This work was published in Pothos and Close (2008).

2.9.2.2 Materials and procedure

As for Experiments 1 – 4, stimuli were circles enclosed in squares, with the circles 'blended in' with the squares (using CorelDraw), so as to make them look more like individual objects (see Figure 11). The similarity structure for the two conditions was again specified on abstract 1 – 10 scales, and these were mapped to the physical dimensions of circle size and square size by assuming a Weber's fraction of 7.5% for both the circles (smallest size: 25 mm) and the squares (smallest size: 50 mm; Morgan, 2005). Each stimulus was printed individually on a piece of paper as large as the stimulus, which was subsequently laminated.

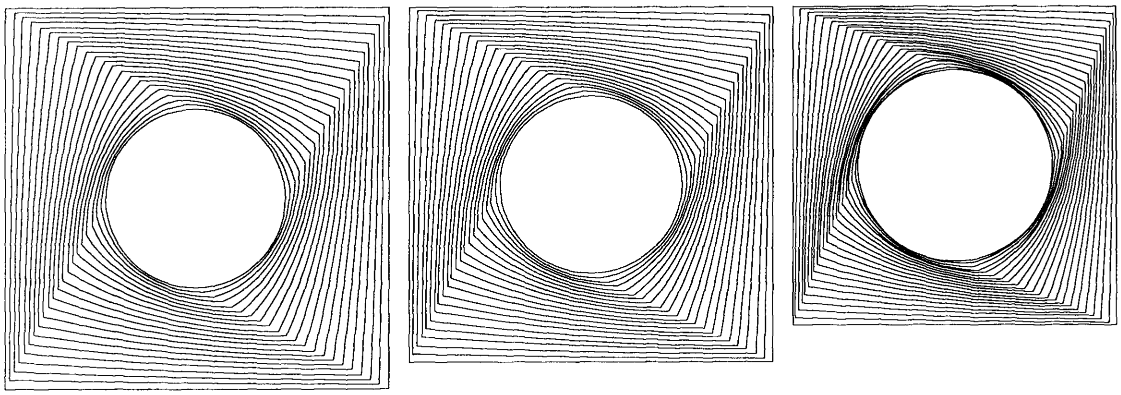

Figure 11. A few examples of the stimuli employed in Experiment 5. The stimulus presented on the left shows the greatest size in the square dimension, and the stimulus presented on the right shows the greatest size in the circle dimension.

As noted, the objective in the present experiment was to create stimulus structures such that Group(*x*)/ Group(*y*) were not superordinate or subordinate relative to Group(*x*,*y*). Figures 12 and 13 show two such structures: Figure 12 shows a stimulus structure for which the simplicity model predicts a preference for unidimensional classification (i.e., Codelength(*x*) and Codelength(*y*) are less than Codelength(*x*,*y*)). Figure 13 shows a stimulus structure for which the simplicity model predicts a preference for two-dimensional classification (i.e., Codelength(*x*,*y*) is less than Codelength(*x*) and Codelength(*y*)). Again, it is important to note that the codelength for the predicted 'optimal' classifications in each condition are approximately the same, as are the codelengths for the predicted 'suboptimal' classifications (of course, in one condition the 'optimal' classification reflects unidimensional classification, and in the other condition the 'optimal' classification reflects two-dimensional classification).

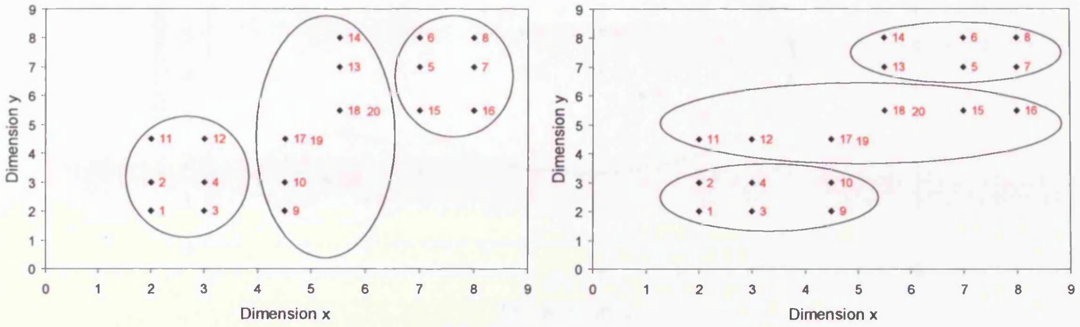

Figure 12. A stimulus structure where the simplicity model predicts a unidimensional classification preference (the unidimensional classifications are shown). Where there are two numbers next to a data point, this means that two identical items were included in the stimulus set. The left-hand structure depicts the most intuitive classification along just dimension *x*, in which the predicted 'optimal' clustering is (1,2,3,4,11,12) (5,6,7,8,15,16) (9,10,13,14,17,18,19,20), and the right-hand structure depicts the most intuitive classification along just dimension *y*, in which the predicted 'optimal' clustering is (1,2,3,4,9,10) (5,6,7,8,13,14) (11,12,15,16,17,18,19,20). Both these classifications are associated with a codelength of 57.6%. When represented along both dimension *x* and *y* together, the predicted 'optimal' clustering is (1,2,3,4,9,10,11,12) (5,6,7,8,13,14,15,16,17,18,19,20), with an associated codelength of 73.4%.

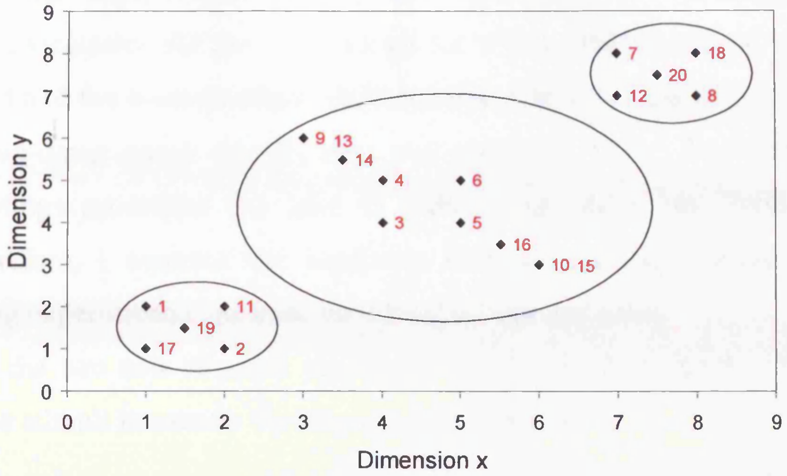

Figure 13. A stimulus structure where the simplicity model predicts a two-dimensional classification preference (the two-dimensional classification is shown). Where there are two numbers next to a data point, this means that two identical items were included in the stimulus set. When represented along both dimension *x* and *y* together, the predicted 'optimal' clustering is (1,2,11,17,19) (3,4,5,6,9,10,13,14,15,16) (7,8,12,18,20), with an associated codelength of 59.4%. The predicted 'optimal' clustering along just dimension *x* is (1,2,3,4,9,11,13,14,17,19) (5,6,7,8,10,12,15,16,18,20) and along just dimension *y* is (1,2,3,5,10,11,15,16,17,19) (4,6,7,8,9,12,13,14,18,20), both with an associated codelength of 73.5%.

Given these new stimulus structures, it is again important to establish that participants perceived the stimuli as I intended. Therefore, in exactly the same way as for Experiment 1, stimulus similarity ratings were collected from 12 participants for each of the stimulus structures (see Section 2.5.1.2 for procedural details). The number of stimuli presented in each condition now totals 20; consequently, participants made a total of 380 similarity comparisons, which reflected rating the similarity of all possible stimulus pairs once, excluding pairs of identical stimuli.

Using these similarity ratings, the Multidimensional Scaling (MDS) procedure derived a spatial representation in two-dimensions for the stimuli. For the stimulus set for which simplicity predicted a unidimensional classification preference, the best solution was associated with a stress of 0.168, and for the stimulus set for which a two-dimensional classification preference was predicted, the best solution was associated with a stress of 0.149. The Orthosim procedure set out in Section 2.5.1.2 was again used to assess the similarity of the MDS-derived representations for the stimuli with the experimenter-assumed coordinates. The similarity coefficient between the coordinates for the stimulus set for which unidimensional classification was predicted and the corresponding MDS solution was 0.74, and for the stimulus set for which two-dimensional classification was predicted, 0.72. To recapitulate, the similarity ratings procedure can lead to rather noisy data (see Section 2.5.1.2). Overall, therefore, I consider the similarity between the MDS solutions and the corresponding experimenter assumed coordinates to be adequate.

With the two new stimulus structures established, participants were asked to categorise the stimuli in exactly the way as in Experiment 1.

2.9.3 Results

The results from Experiment 5 are presented in Figure 14 (see Figure 32 of Appendix 1 for the frequency with which participants produced classifications based on a specific number of clusters). As can be seen, the pattern of results of Experiment 5 is opposite to those of Experiments 1 – 4. As such, the results of Experiment 5 are consistent with the predictions of the simplicity model. That is, in the case where simplicity predicted a preference for unidimensional classification, participants' classifications were more similar to Group(x) and Group(y) than to Group(x,y), and in the case where simplicity predicted a preference for two-dimensional classification, participants' classifications were more similar to Group(x,y) than to Group(x) or Group(y).

Greenhouse-Geisser corrected ANOVA, with condition (predicted unidimensional preference or predicted two-dimensional preference) as a between-participants factors and Rand similarity (Rand(x) or Rand(y) or Rand(x,y)) as a within-participants factor, revealed no effect of condition, $F(1, 38) = 1.21, p > .05$, an effect of Rand similarity, $F(1.43, 54.26) = 14.75, p < .001$, and a significant interaction between these factors, $F(1.43, 54.26) = 68.64, p < .001$. Tests of simple main effects revealed that there was a significant effect of condition at Rand(x), Rand(y), and Rand(x,y) (smallest $F(1, 114) = 4.25, p < .05$). Simple main effects further revealed that there was a significant effect of Rand similarity in the condition where a preference for unidimensional classification was predicted ($F(1.72, 32.76) = 51.16, p < .001$), and in the condition where a preference for two-dimensional classification was predicted ($F(1.09, 20.68) = 33.69, p < .001$).

Critically, in the condition where simplicity predicted a preference for unidimensional classification, the similarity of participants' classifications to Group(*x*,*y*) was significantly less than to both Group(*x*) and Group(*y*) (as assessed with Bonferroni-adjusted paired samples t-tests, $t(19) = -11.06, p < .001$, and, $t(19) = -2.95, p < .009$, respectively). In the condition where simplicity predicted a preference for two-dimensional classification, similarity to Group(*x*,*y*) was significantly greater than to both Group(*x*) and Group(*y*) (assessed in the same way, $t(19) = 21.73, p < .001$, and, $t(19) = 6.44, p < .001$, respectively).

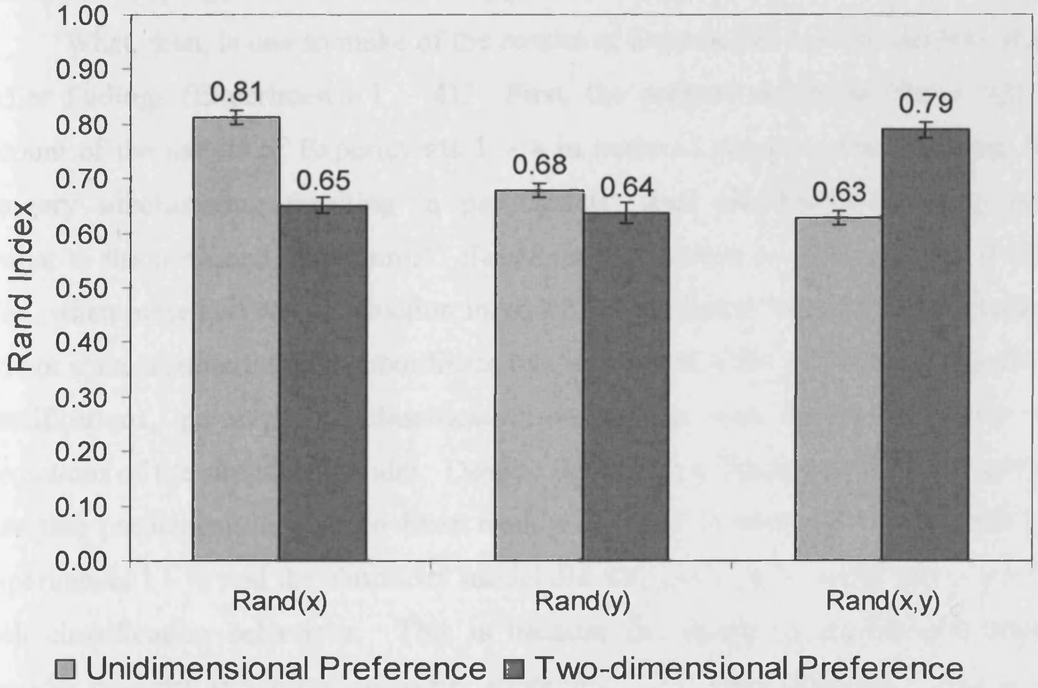

Figure 14. The results of the Rand Index analyses for Experiment 5. Rand(*x*) means the Rand similarity of participants' classifications to Group(*x*), etc. 'Unidimensional Preference' refers to the condition where the simplicity model predicted a preference for unidimensional classification. 'Two-dimensional Preference' refers to the condition where the simplicity model predicted a preference for two-dimensional classification. Error bars denote the standard error.

2.9.4 Discussion

Experiment 5 employed two new stimulus structures in which the predicted 'optimal' classifications did not share a superordinate-subordinate relationship with the predicted 'suboptimal' classifications. For one of these stimulus structures, the simplicity model predicted a unidimensional classification preference, and for the other stimulus structure, simplicity predicted a two-dimensional classification preference. In contrast to the findings of Experiments 1 – 4, participants' classification preferences were found to be entirely consistent with the predictions of the simplicity model. That is, in the case where the simplicity model predicted unidimensional classification, participants' classifications were more similar to Group(*x*) and Group(*y*) than to Group(*x*,*y*), and in the case where the simplicity model predicted two-dimensional classification, participants' classifications were more similar to Group(*x*,*y*) than to either Group(*x*) or Group(*y*).

What, then, is one to make of the results of Experiment 5 in the context of the earlier findings (Experiments 1 – 4)? First, the present results do encourage an account of the results of Experiments 1 – 4 in terms of category subclustering; this category subclustering resulting in participants' final classifications being most similar to the predicted 'suboptimal' classifications in each condition. Why is this? Well, when presented with a situation in which the predicted 'optimal' classifications did not share a superordinate-subordinate relationship with the predicted 'suboptimal' classifications, participants' classification behaviour was found to match the predictions of the simplicity model. Despite the results of Experiment 5, it is still the case that participants appear to have readily engaged in category subclustering in Experiments 1 – 4, and the simplicity model did not, and indeed would never, predict such classification behaviour. This is because the simplicity model will always consider category subclustering to be 'suboptimal' (i.e., less intuitive). This is not surprising; natural categories often have a kind of hierarchical structure – in which the basic level (e.g., dog) shares a superordinate-subordinate relationship with the subordinate level (e.g., Poodle) – and yet, in general, people choose to classify such stimuli at the basic level (see Rosch, et al., 1976). It is likely, therefore, that participants' apparent 'preference' for category subclustering in Experiments 1 – 4 is partly a product of the artificial nature of the experimental task employed (e.g., the use of simple geometric stimuli, etc.). Overall, therefore, it is difficult to speculate about the ecological validity of the simplicity model from the present findings.

In summary, the success of the simplicity model to accurately predict participants' classification behaviour appears to be dependent on whether a predicted 'optimal' classification shares a superordinate-subordinate relationship with the predicted 'suboptimal' classification(s). Before drawing some general conclusions from the findings of Experiments 1 – 5, it is first interesting to enquire whether other models of unsupervised categorisation are better able to capture the general patterns of results found in this chapter (albeit *post hoc*).

2.10 Other models

First, supervised models of categorisation, which employ free parameters for attentional weighting, would likely be able to describe all of the results of Experiments 1 – 5. However, it is unclear whether such models would predict these results without some constraints on determining these free parameters *a priori* (e.g., Nosofsky, 1989). With respect to the models of unsupervised categorisation outlined earlier, it was noted that SUSTAIN spontaneously classifies a set of stimuli on the basis of more than one dimension when (and for) dimensions that are highly intercorrelated with each other. Focusing first on the stimulus structures employed in Experiments 1 – 4, SUSTAIN would appear to predict the following: for the stimulus structure of Figure 5, where simplicity predicted a preference for unidimensional classification, the correlation between the two dimensions of variation was found to be -.002, ($p > .05$). For the stimulus structure of Figure 6, where simplicity predicted a preference for two-dimensional classification, the correlation between the two dimensions of variation was found to be .457 ($p > .05$). Broadly, therefore, these correlations suggest a similar pattern of predictions to the simplicity model, indicating a strong unidimensional classification bias for Figure 5, and a tendency towards a two-dimensional classification bias for Figure 6. Of course, these predictions were not supported. While the correlation between dimensions x and y is not particularly high in the case where simplicity predicted a two-dimensional classification preference (Figure 6), it is at least substantially higher than the correlation between dimensions x and y in the case where simplicity predicted a preference for unidimensional classification.

Based on the assumptions of SUSTAIN, therefore, the finding that participants showed an overall preference for unidimensional classification in the case where a preference for two-dimensional classification was predicted is not all that surprising. However, based on the same assumptions, the finding that participants showed a preference for two-dimensional classification in the case where a preference for unidimensional classification was predicted is highly surprising (to reiterate, there existed almost zero correlation between dimensions *x* and *y* for the stimulus structure of Figure 5).

Focusing now on Experiment 5, for the stimulus structure of Figure 11, where simplicity predicted a unidimensional classification preference, the correlation between dimensions *x* and *y* was .763 (p<.01). For the stimulus structure of Figure 12, where simplicity predicted a two-dimensional classification preference, the correlation between dimensions *x* and *y* was almost identical (.760, p<.01). It is apparent, therefore, that SUSTAIN does not specify the unidimensional versus two-dimensional bias that was derived from the simplicity model; indeed, one can infer from these correlations that SUSTAIN would predict a preference for two-dimensional classification in both cases. However, while the simplicity model was specifically constructed to deal with classification based on the simultaneous presentation of stimuli, SUSTAIN is an incremental model of category learning. Consequently, it is possible that the conclusions just drawn may be somewhat unfair to SUSTAIN (although on average, any such discrimination against SUSTAIN should be ruled out).

Can the results of Experiments 1 – 5 be captured by the statistical clustering algorithms discussed briefly in Chapter 1? A couple of points are important to consider here: first, many statistical algorithms are not suitable here as they do not have a ready psychological interpretation. Second, while certain versions of *K*-means clustering can be considered, as these algorithms specify clustering by maximising within-cluster similarity while minimising between-cluster similarity (similar to the simplicity model), as discussed in Chapter 1, they require information to be given about the number of categories sought (*K*). As highlighted throughout Chapter 1 and this chapter, when assessing participants' preference for unidimensional versus multidimensional classification, such information may prejudice the issue (see Murphy, 2002).

2.11 General Discussion

When asked to classify a set of stimuli without any feedback, participants will readily engage in this task. Intriguingly, the majority of laboratory research on human unsupervised categorisation has documented an overwhelming and robust bias for unidimensional classification (e.g., Ashby et al., 1999; Medin et al., 1987; Regehr & Brooks, 1995). This unidimensional classification preference is, however, inconsistent with our current understanding of real world categorisation (e.g., Rosch & Mervis, 1975). While multidimensional (family resemblance) classification has been documented in the laboratory, these observations have often only been made after employing a specific task manipulation: this has included manipulations of stimulus format (e.g., Milton & Wills, 2004), procedural details (e.g., Milton et al., 2008), and the introduction of prior knowledge (e.g., Ahn, 1990, 1991; Kaplan & Murphy, 1999; see also, Medin et al., 1987). The work presented in this chapter highlights the critical importance of stimulus similarity structure in influencing human unsupervised categorisation. Specifically, I have shown that, like SUSTAIN (Love et al., 2004), the simplicity model of unsupervised categorisation also predicts a unidimensional classification preference for the binary stimulus structure of Medin et al. (1987; see Figure 1, Chapter 1), on the basis of its abstract stimulus structure. To reiterate, this is consistent with Medin et al.'s (1987) findings. Critically, Experiment 5 of this chapter documented the first empirical demonstration of a preference for two-dimensional classification, based solely on a set of stimuli's abstract similarity structure.

To assess unidimensional versus two-dimensional classification, I employed the simplicity model of unsupervised categorisation (Pothos & Chater, 2002) to derive a number of classification predictions about two stimulus structures, and the Rand Index analysis to compare participants' classifications with the predicted classifications. For all experiments, one stimulus structure was derived where simplicity predicted a preference for unidimensional classification, and one stimulus structure was derived where simplicity predicted a preference for two-dimensional classification. In Experiments 1 – 4, the pattern of results found did not support the predictions of the simplicity model; in fact, the results were in the opposite direction to the model's predictions. These results appeared to reflect an instance of category subclustering. Indeed, focusing on the findings of Experiments 4 and 5 together, this suggested that participants had a preference for this kind of classification. However, due to the fact that the predicted 'optimal' classifications shared a superordinate-subordinate relationship with the predicted 'suboptimal' classifications in Experiments 1 – 4, it was not possible to unambiguously confirm a unidimensional versus multidimensional bias using the Rand Index. Consequently, in Experiment 5, two new stimulus structures were derived in which the predicted 'optimal' classifications did not share a superordinate-subordinate relationship with the predicted 'suboptimal' classifications. By employing these two new stimulus structures, participants' classification behaviour was found to be consistent with the predictions of the simplicity model.

The results of this chapter, therefore, are important in informing our understanding of why participants often show a strong bias for unidimensional classification in the laboratory. To recapitulate, the unidimensional classification bias documented in many previous studies of unsupervised categorisation is odd given the nature of our everyday categories, which are based on a principle of family resemblance (Rosch & Mervis, 1975; Wittgenstein, 1953). One likely explanation for this laboratory-based unidimensional unsupervised categorisation bias, therefore, is that it is simply an artefact of the experimental procedures that have been employed. The work presented in this chapter supports this claim, and suggests that this artefact likely stems from a lack of understanding about the biases that are inherent within the similarity structure of the stimuli employed (e.g., Medin et al., 1987). For example, based on the binary stimulus structure of Medin et al. (1987), unidimensional classification should be considered more intuitive.

What do the contrasting findings of Experiments 1 – 4 and Experiment 5 mean for the validity of the simplicity model? First, they indicate that the model is only accurate in its predictions when dealing with stimulus structures where the basic level of classification does not have obvious substructure. This is a bit of a problem for the simplicity model, as many category structures have some sort of hierarchical structure. It is interesting that with respect to everyday categories, however, people often prefer basic level categorisation over subordinate level categorisation (Rosch et al., 1976). As noted earlier, it seems likely that participants' tendency to engage in any category subclustering may have been driven by the artificial nature of the experimental task. The fact is this though, in Experiments 1 – 4, participants' classification behaviour simply did not match the classifications that were predicted to be 'optimal' (more intuitive) by the simplicity model. Whether this was because participants actually preferred unidimensional classification when a two-dimensional classification preference was predicted, for example, or it was brought about through category subclustering, the above fact is clearly a major limitation of the simplicity model. In its favour, of course, is the fact that the predictions of the simplicity model were supported in Experiment 5.

Interestingly, it is apparent that SUSTAIN (Love et al., 2004), for example, also fails to successfully capture the results of Experiments 1 – 4. Moreover, SUSTAIN further appears to fail to capture the results of Experiment 5 (on the basis of the correlations that exist between the dimensions of variation, at least). Of course, there are other model approaches to unsupervised categorisation which have not been considered here (e.g., Compton & Logan, 1993; Schyns, 1991). A number of considerations guided the emphasis on the simplicity model and SUSTAIN (and also on the Rational model earlier in the chapter). With respect to modelling human cognition, compelling arguments have recently been made for the relevance of simplicity and Bayesian principles in this endeavour (e.g., Chater, 1999; Feldman, 2000; Tenenbaum, Griffiths, & Kemp, 2006). These models are proven in terms of their flexibility to capture a whole range of unsupervised categorisation data. Furthermore, the free parameters employed in both the Rational model and SUSTAIN have commonly been fixed over the course of various model demonstrations. Perhaps most important of all, however, is that the unidimensional bias documented in many unsupervised categorisation experiments has previously been directly investigated using SUSTAIN (Love et al., 2004).

In conclusion, the results of Chapter 2 demonstrate that stimulus similarity structure influences people's classification behaviour. Critically, the findings of Experiment 5 of this chapter document the first empirical observation of a preference for multidimensional unsupervised categorisation, on the basis of the abstract stimulus structure of the stimuli. However, while similarity structure is clearly influential, the results of Experiments 1 – 4 reinforce the sense that human classification is a complex phenomenon, driven by factors that transcend pure perceptual similarity. This point is important, and it highlights the clear limitations of the simplicity model. Indeed, with respect to natural, everyday categorisation, the simplicity model is limited in a number of ways: First, due to the combinatorics of the simplicity model, classification that has to take into account many different stimulus dimensions, and many thousands of stimuli, would require a vast amount of computational power. Second, the simplicity model was specifically developed to model simultaneous unsupervised categorisation. However, in the real-world, stimulus classification will most often be sequential, occurring over a period of time. Consequently, stimulus categorisation will involve a memory component. Furthermore, while the simplicity model is able to find structure within a set of stimuli and form categories according to this, it is rather inflexible. By contrast, SUSTAIN, for example, was developed to be highly flexible. As Love et al. note, "the categorisation system must be able to both assimilate structure and discover or even create that structure" (2004, p. 309). One critical mechanism for determining category structure in SUSTAIN, at least, is 'surprisingness'. That is, if a novel stimulus is sufficiently surprising (i.e., it exceeds some threshold level of dissimilarity to an already formed category), then this is a good indicator that SUSTAIN should create a new category in which to accommodate the novel stimulus. Importantly, this parameter is flexible, based on prior stimulus experience. Indeed, 'surprisingness' has been shown to be an important mechanism for unsupervised category construction by Clapper and Bower (1994, 2002).

In Chapter 3 of this thesis, I was interested in moving away from the specific question of unidimensional versus multidimensional unsupervised classification. Specifically, I wanted to explore aspects of unsupervised categorisation that are beyond the scope of the simplicity model. This included assessing the influence of 'surprise', as well as other features of perceptual experience, which might influence whether stimuli are 'classified together' or 'classified apart'. Moreover, I was keen to assess unsupervised categorisation using a procedure and stimuli that would more accurately reflect natural unsupervised categorisation (i.e., by sequentially exposing people to naturalistic stimuli that are composed of many dimensions of variation).

Chapter 3

Within-category similarity structure and incidental unsupervised categorisation

3. Introduction

The experiments of Chapter 2 highlighted the importance of abstract similarity structure in influencing people's unsupervised classification behaviour. That is, people were shown to be sensitive to the perceived regularities and discontinuities (or similarity-based relationships) that exist within a stimulus set (albeit not in a manner that is always consistent with the predictions of the simplicity model; Pothos & Chater, 2002). Consequently, while certain stimulus structures were found to support classification based on a single dimension of variation (e.g., dimension *x*), other stimulus structures were found to support classification based on more than one dimension of variation (e.g., dimension *x*,*y*). However, as has been highlighted throughout this thesis, our 'natural', everyday categories are not unidimensional in kind. Rather, research has clearly established that our everyday categories are rich, broad constructs, based on a principle of family resemblance (Rosch, 1973, 1975; see also, Wittgenstein, 1953). In Rosch's (1973, 1975) terms, natural categories have an *internal structure*; consequently, not all items are equally good members of a category.

Numerous differences exist between unsupervised categorisation that occurs naturally and that performed by participants in the experiments of Chapter 2 of this thesis. In natural unsupervised categorisation, category formation is incidental (Clapper & Bower, 1994; Love, 2002): this requires a person to a) realise that there is structure present, and b) to then utilise this structure to guide their classifications. In contrast, in the experiments of Chapter 2 of this thesis (and in the majority of previous investigations of unsupervised categorisation), participants were explicitly told to categorise a set of stimuli. This explicit instruction to categorise, therefore, will likely promote a belief in participants that their task is to find some experimenter defined category structure (i.e., the category structure that makes the most 'intuitive' sense)[8]. Consequently, rather than identifying category structure incidentally, participants will be intentionally seeking structure within the experimental materials. To recapitulate from Chapter 0, this is important because these different forms of categorisation (intentional versus incidental) have been associated with different kinds of unsupervised classification. That is, while intentional unsupervised categorisation has been associated with more 'rule-like' (unidimensional) category learning, incidental unsupervised categorisation has been associated with classification based on family-resemblance (Love, 2002; the latter reflecting categorisation that is compatible with the nature of our everyday categories). Moreover, natural unsupervised categorisation will rarely, if ever, proceed under conditions of simultaneous stimulus exposure, as occurred in the experiments of Chapter 2. Instead, stimulus exposure will most likely be sequential, and stimulus categorisation will therefore involve a memory component: that is, stimulus comparisons will not be made on the basis of their veridical physical dimensions, but rather on participants' stored representations of those stimuli (Clapper & Bower, 2002; Love et al., 2004). Finally, natural unsupervised categorisation will proceed with respect to complex stimuli constructed from many different dimensions of variation, rather than simple stimuli constructed from just a couple of dimensions of variation. Consequently, it is far harder to identify some defining feature for complex naturalistic stimuli compared to simple artificial stimuli. All these factors, therefore, will play a role in determining that natural, everyday categories reflect a principle of family resemblance.

The experiments reported in this chapter introduce a broader approach to the study of unsupervised categorisation. Specifically, these experiments focus on incidental unsupervised categorisation, following the sequential presentation of complex stimuli: for the purposes of this thesis, I will simply refer to this kind of unsupervised categorisation as *incidental categorisation.* The main benefits of studying incidental categorisation are two-fold: First, it affords a more naturalistic approach to the investigation of unsupervised categorisation. Second, it affords the unique ability to assess unsupervised categorisation in nonlinguistic agents. Clearly, it is not possible to ask a nonhuman animal to classify a set of stimuli in a way that

[8] Even if participants are told to group a set of stimuli in a natural and intuitive way, and that there is no correct answer, given the experimental situation they are in, why should they believe this?

feels 'natural and intuitive to them'. However, the procedures detailed below set out one way in which it is possible to investigate whether nonhuman animals, like humans, *spontaneously* group together stimuli in any meaningful sense. Moreover, in both humans and nonhuman animals, the procedures detailed below allow for an assessment of the conditions that may or may not promote the spontaneous classification together, or spontaneous classification apart, of different stimuli. Specifically, the experiments reported in this chapter sought to assess how within-category similarity structure influences the incidental classification of similar, but distinct stimuli in both humans and rats. In taking this comparative approach, I hope to assess more fully the role of the classifier in unsupervised categorisation.

3.1 Investigating incidental categorisation

To investigate incidental categorisation, the experiments reported in this chapter exploit a well-known influence that categorisation can have over the phenomenon of stimulus generalisation. The first formal demonstration of stimulus generalisation was described by Pavlov (1927), who observed that once a dog had come to show a conditioned salivary response to a tone of a specific frequency, other tones that were close in frequency to the trained tone would also "spontaneously" provoke salivation. The close relationship between stimulus generalisation and similarity has been widely documented (see Pearce, 1994). Shepard (1987) has shown that stimulus generalisation follows a lawful relationship with similarity, such that the amount of generalisation between two stimuli decays exponentially with their decreasing similarity. Interestingly, a number of authors have proposed that categorisation warps psychological similarity space (Nosofsky, 1989), such that categorisation influences the perceived similarity between stimuli. For example, Livingston, Andrews and Harnad (1998; see also, Kurtz, 1996) have shown that if participants are taught that stimuli are members of the same category, then they will later perceive these stimuli to be more similar than participants that did not learn this classification. The reverse is also true; if participants are taught that stimuli are members of contrasting categories, then they will later perceive these stimuli to more distinct than participants that did not learn this classification (Goldstone, 1994). The two most commonly used terms to describe these compression and expansion effects are categorical perception (Harnad, 1987), and acquired equivalence and distinctiveness (Goldstone, 1998; Hall, 1991; Lawrence, 1949).

The classic finding from research on categorical perception is that participants find it harder to distinguish between physically different stimuli when they come from the same category than when they come from different categories (see Harnad, 1987). For example, within the domain of speech perception, Liberman, Harris, Eimas, Lisker and Bastian (1957) found that participants were more accurate to confirm that a sound X was identical either to a sound A or a sound B when syllables A and B belonged to different phonemic categories than when they were variants of the same phoneme (the physical differences between A and B were equated between conditions). In their task, the three sounds were presented sequentially to participants (i.e., A followed by B followed by X). While the majority of research in this area has focused on colour categories (Bornstein, 1987) and phoneme categories (e.g., Pastore, 1987), categorical perception effects have also been shown to occur with other visual stimuli (e.g., Livingston et al., 1998). Newell and Bülthoff (2002), for example, morphed together naturalistic stimuli from within the same basic level category (e.g., Wine-Coke bottle) and from different basic level categories (e.g., Bottle-Lamp). Using this morphing technique, they rendered 11 object images from each morph continuum. Initially, participants engaged in an 'XAB' discrimination task, in which they were presented with one stimulus, X, followed by the simultaneous presentation of two other stimuli, A and B. Stimuli A and B always differed from each other in their physical appearance, and stimulus X was identical to either stimulus A or B. Participants' task was to decide whether stimulus X was identical to stimulus A or B. Subsequently, participants engaged in an identification task, in which they were asked to classify each morph image as either one end of a morph continuum (e.g., Wine bottle) or the other end of a morph continuum (e.g., Coke bottle). Employing a commonly used technique to index categorical perception (see Calder, Young, Perrett, Etcoff, & Rowland, 1996) – in which participants' discrimination performance is predicted from the identification data, based on the assumption that the objects were categorically perceived – Newell and Bülthoff (2002) reported categorical perception for all object pairs created by morphing together two objects from the same basic level category. In a final experiment, inter-object perceptual similarity was shown to be closely correlated with categorical perception; the greater the similarity between a set of objects, the more likely it is that they will be perceived categorically (Newell & Bülthoff, 2002).

While learned categorical perception has been widely documented through the use of supervised training procedures, as noted by Gureckis and Goldstone, "it remains a somewhat opaque question if learned CP [categorical perception] effects are restricted to cases where subjects make a differential response to each category or if other aspects of category organisation, such as the similarity structure or distribution of items within a category, may also exert an influence on perception" (2008, p. 1876). In a recent conference paper, however, Gureckis and Goldstone (2008) reported some preliminary evidence for an unsupervised categorical perception effect, concluding that participants *are* sensitive to sources of within-category structure, as predicted by the SUSTAIN model of category learning (Love et al., 2004).

Learned discrimination-based studies clearly document that stimulus similarity is influenced by their classificatory status (or shared associative history). Given the relationship between similarity and stimulus generalisation (see Pavlov, 1927; Shepard, 1987), a number of studies, concentrated mainly within the domain of animal learning, have not surprisingly also shown that stimulus generalisation is directly influenced by the classificatory status of a set of stimuli (that is, whether the stimuli have acquired equivalence or distinctiveness). For example, Honey and Watt (1998, 1999; see also, Honey & Hall, 1989) initially gave rats training in a conditional discrimination task in which stimuli A and B (but not C and D) signalled a food reward when presented with a cue X, and stimuli C and D (but not A and B) signalled a food reward when presented with a cue Y. Following discrimination training, in which A and B, and, C and D should have acquired equivalence, stimulus A, but not C, was paired with a mild footshock. Subsequently, stimulus B was found to elicit a greater fear response than stimulus D. That is, the discrimination training employed by Honey and Watt altered the effective similarity of stimuli A, B, C and D, such that stimuli A and B came to be perceived as more similar than stimuli A and D. Similar patterns of generalisation behaviour have been reported in humans by Hodder, George, Killcross, and Honey (2003). In one of their experiments, participants were taught a conditional discrimination in which they learned that a person would suffer an allergic reaction if they ate meat products A and B (but not C and D) with vegetable X, and if they ate meat products C and D (but not A and B) with vegetable Y. Interleaved between the previous discrimination training, participants either received training in a second congruous or incongruous conditional discrimination involving two further vegetables (V and W). In condition Congruous, participants learned that a person would suffer an allergic reaction if they ate meat products A and B (but not C and D) with vegetable V, and if they ate meat products C and D (but not A and B) with vegetable W. In condition Incongruous, participants learned that a person would suffer an allergic reaction if they ate meat products A and D (but not B and C) with vegetable V, and if they ate meat products B and C (but not A and D) with vegetable W. In general, Hodder et al. (2003) found that the initial conditional discrimination training generalised better to participants in condition Congruous than it did to participants in condition Incongruous.

In summary, there exists a large body of evidence using supervised training procedures to support the claim that categorisation alters the effective similarity of stimuli, which directly influences the amount of subsequent stimulus generalisation. However, research that has directly addressed the influence that unsupervised categorisation has on altering the effective similarity of stimuli is extremely limited. Moreover, there has been little investigation into how within-category similarity structure influences the incidental classification of stimuli, and subsequent stimulus generalisation. The experiments reported in this chapter, therefore, sought to directly assess how within-category similarity structure influences incidental stimulus categorisation. Specifically, I was interested in assessing what aspects of within-category similarity structure influence whether stimuli are spontaneously classified together, or spontaneously classified apart, in both humans and rats. In the following sections, therefore, I outline a number of factors that may be influential in determining the incidental classification of stimuli.

3.1.1 Transformational knowledge

Objects in the environment can be seen to have a 'natural' direction; that is, they evolve in a manner that is principled (Hahn, Close, & Graf, 2009; Zaki & Homa, 1999). For example, a tadpole has to undergo a marked change before it becomes a frog, and...

> "If one were to look at these entities separately, having no knowledge of the nature of these changes, one might find it difficult to classify them as belonging to the same category. However, given the intermediate steps between the tadpole and the frog, it becomes easier to identify the two examples as being forms of the same category".
>
> (Zaki & Homa, 1999, p. 70)

The appreciation of the various steps that lie intermediate within the transformation of one object into a different object has been termed *transformational knowledge* (Zaki & Homa, 1999). Based on the aforementioned reasoning, Zaki and Homa (1999) proposed that the acquisition of an object concept – one's mental understanding of what constitutes a member of a specific category – will be facilitated by exposure to that object's successive changes. Over four experiments using dot patterns, Zaki and Homa (1999) reported evidence to support their view. Transformational knowledge was found to enhance category learning, and more specifically, participants' classification of novel items was shown to be better following category training that progressed in a systematic order rather than in a random order. They further found that novel patterns that lie on the transformational path were categorised more quickly only when participants had previously received systematic category training. Their results are consistent with research in faces that has shown that people are able to recognise faces that have undergone dynamic change (e.g., Seamon, 1982), and also with the phenomenon of representational momentum (e.g., Freyd & Finke, 1984). For example, Freyd and Finke (1984) found that when participants experienced a series of displays that implied rotation in a presented pattern, their short-term visual memory for the final position of that pattern was shifted forward along the direction of implied movement. Finke, Freyd and Shyi argued that "the induced shifts in visual memory occur because there is a natural tendency to mentally extrapolate implied motions into the future" (1986, p. 176).

It seems plausible, therefore, to suppose that transformational knowledge may play an important role in encouraging the incidental classification of two different stimuli into the same category. However, based on the findings of Newell and Bülthoff (2002) reported above – who showed that participants often perceive a set of morphed stimuli categorically by imposing a category boundary at some point along the morph continuum – it is possible that the presence of transformation knowledge between two stimuli may actually serve to reinforce the sense that the two different stimuli are distinct, and should therefore be classified apart into different categories. Consequently, this may mean that the two stimuli are actually treated as more different from each other than if transformational knowledge had not been present. Of course, it is also possible that these two influences may cancel each other out, leaving participants unsure about the classificatory status of the two different stimuli.

3.1.2 Surprise-driven category invention

While some theories of spontaneous category learning have proposed that learning operates through explicit iterative hypothesis testing (Billman & Knutson, 1996), other theories have proposed that, within a stimulus domain, correlational patterns are captured and are used to explicitly partition the stimuli (Clapper & Bower, 1994, 2002; Love et al., 2004). These theories link the formation of new categories (or clusters) to unexpected changes in stimulus structure, which creates surprise within the categoriser (e.g., Clapper & Bower, 2002; Love et al. 2004).

Specifically, Clapper and Bower (1994) proposed that a novel exemplar is compared with respect to a person's normative expectations (summary knowledge) about what it means to be a member of, for example, Category A. If this novel exemplar fits poorly into Category A, then it is likely that a new category will be invented to accommodate this distinct exemplar. Whether or not a new category will be invented depends on the level of surprise generated on presentation of the novel exemplar; specifically, "the probability of creating a new category in response to the first instance of Category B should increase with the number (*n*) of prior instances of Category A" (Clapper & Bower 1994, p. 447). For example, following a single presentation of an instance of Category A, only a weak set of norms will have been established determining Category A membership. This means that presentation of a Category B instance will not make for a particularly surprising contrast to the Category A norms, and will not, therefore, warrant the invention of a new category. However, following many presentations of instances of Category A, a strong set of norms will have been established determining Category A membership. Consequently, when an instance of Category B is presented, it will readily violate these well-established norms, which would constitute a surprising stimulus event. In this case, therefore, the Category B instance will most likely be accommodated in a newly invented category. Using an attribute-listing task, Clapper and Bower (1994, 2002) found good support for a surprise-driven category invention mechanism in unsupervised categorisation. Specifically, they found that participants were more likely to engage in category invention following blocked stimulus presentation compared to intermixed stimulus presentation. One implication of this is that participants should come to perceive a set of stimuli as less similar following blocked exposure than following intermixed exposure. Interestingly, this conclusion sits in contrast to formally equivalent results in nonhuman animals, detailed in Section 3.1.3. Briefly, a common finding in nonhuman animals is that two stimuli will be perceived as less similar following intermixed exposure (i.e., AX-BX-AX-BX) than blocked exposure (i.e., AX-AX-BX-BX; e.g., Symonds & Hall, 1995; Honey, Bateson & Horn, 1994). Equivalent findings have also been reported in humans (e.g., Lavis & Mitchell, 2006; Dwyer, Hodder & Honey, 2004). This suggests, therefore, that the contrasting results with those of Clapper and Bower (1994, 2002) are possibly due to the operation of different processes (i.e., "feature detection" as opposed to stimulus classification; Gibson, 1963), brought about through procedural and stimulus differences. Indeed, one particularly salient difference is the nature of the stimuli employed: while human and nonhuman animal studies investigating perceptual learning have typically employed just a couple of non-variable stimuli, participants in Clapper and Bower's (1994) study were presented with many complex stimuli, containing much variability. It is possible, therefore, that if one were to introduce a greater degree of stimulus variability in investigations of the intermixed versus blocked effect in perceptual learning, one might see the greatest reduction in stimulus similarity following blocked exposure, due to stimulus classification.

In SUSTAIN, a cluster can represent individual stimulus exemplars, a subset of feature values within a category, or the representation of a category as a whole (Love et al., 2004). With respect to unsupervised category construction, the notion of surprise again plays an important role in determining when SUSTAIN creates a new cluster (category). Here, 'surprisingness' reflects dissimilarity: that is, if a stimulus is sufficiently dissimilar from a previous cluster (and therefore makes for a sufficiently surprising event on presentation), SUSTAIN will recruit a new cluster (category) to house that stimulus. Two factors, therefore, influence how surprising a stimulus is: First, there is the similarity of the novel instance to existing clusters. Second, there is the threshold level of dissimilarity required before a new cluster will be recruited. The lower this threshold, the more surprising a moderately dissimilar instance to existing clusters will be perceived (at least, that is the assumption). If little variability exists in certain critical stimulus dimensions which have been selectively attended to, then smaller differences on these dimensions will be regarded as surprising. Consequently, as for the category invention mechanism, the more Category A exemplars that are presented before a Category B exemplar is presented, the more likely it is that a new cluster will be recruited to accommodate the Category B exemplar (Gureckis & Love, 2003). This is because increased exposure to a set of stimuli that are likely members of the same category will allow for the attentional mechanism of SUSTAIN to become tuned to the variability that exists in these stimuli. Consequently, a lower level of dissimilarity will be required for SUSTAIN to recruit a new cluster.

In summary, both the category invention mechanism and SUSTAIN suggest that a novel stimulus should recruit a new category (cluster) if the stimulus constitutes a change in stimulus structure of sufficient magnitude (i.e., it makes for a surprising enough event).

3.1.3 Stimulus exposure and stimulus similarity

While categorisation is one way in which the effective similarity of stimuli can be altered, it is important to note that mere exposure to stimuli, in the absence of any obvious spontaneous categorisation, has also been found to influence stimulus similarity. One phenomenon in this context has been termed *perceptual learning*, which involves "relatively long-lasting changes to an organism's perceptual system that improve its ability to respond to its environment and are caused by this environment" (Goldstone, 1998, p. 586). That is, exposing humans and nonhuman animals to two similar stimuli (e.g., AX and BX) often results in a decrease in their effective similarity to each other, such that later discrimination between these stimuli is facilitated and generalisation between them is reduced (see Gibson, 1963, 1969; Goldstone, 1998; Hall, 1991; McLaren & Mackintosh, 2000). Whether or not an effect of perceptual learning is shown has been found to be influenced by a number of different factors. One particularly important factor in this regard concerns the temporal dynamics of stimulus preexposure.

For example, in rats, Symonds and Hall (1995; see also, Honey et al., 1994) found that intermixed preexposure to two flavour compounds (i.e., AX-BX-AX-BX) resulted in less generalisation of a later conditioned aversion from AX to BX than in a condition in which rats received blocked preexposure to the same stimuli (i.e., AX-AX-BX-BX). Using chequerboard patterns, Lavis and Mitchell (2006; see also, Dwyer et al., 2004) have shown equivalent results to those of Symonds and Hall (1995) in humans. Lavis and Mitchell (2006) further found that participants were more accurate to respond that two chequerboard patterns were different following intermixed preexposure than following blocked preexposure. In this case, then, it is apparent that when two similar stimuli are presented more closely in time, perceptual learning is enhanced (this is in line with the predictions of Gibson, 1969)[9]. Interestingly, Bennett and Mackintosh (1999) have, however, shown that if the time between intermixed preexposure of stimuli AX and BX is reduced to some nominal value just above zero seconds, then rats will actually come to show increased levels of generalisation between AX and BX, relative to other rats that received intermixed stimulus preexposure that incorporated a short temporal delay between presentations of the two stimuli. This result confirmed an earlier finding by Honey and Bateson (1996): they found that later discrimination learning in chicks was worse following intermixed preexposure that incorporated a short interval (mean: 14 sec) between stimulus presentations compared to intermixed preexposure that incorporated a longer interval (mean: 28 sec) between stimulus presentations. This increase in the effective similarity of stimuli following mere exposure can be considered an instance of sensory preconditioning (see Hall, 1991).

When stimulus exposure will lead to perceptual learning and when it will lead to sensory preconditioning has proved notoriously difficult to predict. For example, Bateson and Chantrey (1972) found that monkeys showed poorer discrimination learning between two stimuli following simultaneous preexposure to these stimuli. Specifically, they simultaneously exposed rhesus monkeys either to the numbers 2 and 5, or, 6 and 8 for 50 days. Following this exposure phase, monkeys were trained to discriminate between the numbers 2 and 5, by rewarding a touch of the number 2 but not of the number 5. Monkeys that had previously been exposed to the numbers 2 and 5 learned this discrimination more slowly than monkeys that had been preexposed to the numbers 6 and 8. This finding was replicated again in monkeys using letters as stimuli, and also in chicks. Interestingly, this was not the finding expected based on the theorising of Gibson (1969). Rather, she argued that perceptual learning should be at its most influential following simultaneous stimulus exposure, as this form of exposure would afford the best chance to compare the similar stimuli. In support of her claim, however, Mundy, Honey and Dwyer (2009; see also, Mundy, Honey & Dwyer, 2007) have recently shown that simultaneous preexposure to two highly similar chequerboard stimuli (e.g., AX-BX, BX-AX) enhanced their later

[9] Of course, as well as meaning that the different stimuli will be presented more closely in time, the intermixed preexposure schedule also affords a greater number of comparisons between the two stimuli.

discrimination more than did preexposure to highly similar chequerboard stimuli in a successive manner (CY-CY, DY-DY). Why their results contradict those of Bateson and Chantrey (1972) is still to be resolved; however, I would argue that one probable cause for this discrepancy lies in the nature of the stimuli used. That is, while the stimuli used by Bateson and Chantrey were readily discriminable to start with, the stimuli used by Mundy et al. (2007, 2009) were not (but, see Gibson & Walk, 1956).

In conclusion, mere exposure to stimuli can influence the effective similarity of stimuli. While the classification of stimuli into the same category or different categories has been proposed to explain some of these results (see Bateson & Chantrey, 1972), more commonly an explanation has been sought with respect to the influences of habituation and latent inhibition, and other associative processes (see Hall, 1991; McLaren & Mackintosh, 2000). With respect to the 'standard' perceptual learning findings from intermixed versus blocked stimulus exposure, it is interesting to note that they are broadly inconsistent with the predictions of SUSTAIN (described above; see Section 3.1.2). That is, while the category learning model of SUSTAIN appears to predict a greater reduction in stimulus similarity following blocked stimulus exposure (due to more effective stimulus classification), intermixed stimulus exposure has traditionally been found to reduce stimulus similarity to a greater extent in nonhuman animals (Honey et al., 1994). As highlighted in Section 3.1.2, however, this may have a lot to do with a lack of stimulus variability in most perceptual learning studies in nonhuman animals.

3.1.4 Conclusions

In summary, categorisation can alter the effective similarity of stimuli. While the majority of evidence for this influence has come from supervised training procedures, there is some preliminary evidence that unsupervised categorisation, based on within-category similarity structure, can also influence stimulus similarity (Gureckis & Goldstone, 2008). However, the evidence for this modulation of similarity through unsupervised categorisation is clearly limited. What is more, to the best of my knowledge, no empirical research exists that has directly compared how different distributions of stimuli within a category affect how these stimuli are incidentally categorised, and how this impacts on later stimulus generalisation. That is, are there certain distributions that encourage the spontaneous classification of stimuli into the same category, while other distributions encourage the spontaneous classification of stimuli into different categories? Moreover, the discrimination based studies that have indexed an influence of categorisation on stimulus similarity have typically used designs in which participants engage in hundreds of experimental trials. However, it seems reasonable to assume that people's sensitivity to category structure (if sufficiently obvious) should be immediate, and that the incidental categorisation of stimuli should be a rapid process that can proceed under conditions of minimal stimulus exposure.

In the present experiments, therefore, I was interested in establishing how within-category similarity structure (i.e., the distributional properties of stimuli within a category) influences incidental categorisation under conditions of minimal stimulus exposure. Based on the research outlined above, it is assumed throughout that, relative to a baseline, the incidental classification of stimuli into the same category will increase later generalisation between these stimuli. In contrast, the incidental classification of stimuli into different categories will decrease later generalisation between these stimuli.

3.2 Experiment 6

3.2.1 Introduction

Experiment 6 compared the influence of different conditions of one-shot stimulus preexposure on later stimulus generalisation in humans. Specifically, participants were allocated to one of four preexposure conditions where they received differential exposure to a set of morph stimuli. These stimuli were created by morphing together two naturalistic objects from the same basic level category (e.g., bird) and then rendering the 1%, 20%, 40%, 60%, 80% and 100% morph images (henceforth labelled A, B, C, D, E and F, respectively). The four preexposure conditions included one baseline condition (Baseline), one surprise condition (Surprise), one systematic transformation condition (Sys_trans), and one scrambled transformation condition (Scram_trans; see Table 1 for details). Specifically, in the Baseline condition, participants received preexposure only to the endpoints of each morph continuum (i.e., stimulus A and stimulus F). Participants in the Surprise condition were preexposed to three highly similar stimuli, taken from one end of the morph continuum, and one distinct stimulus that represented the most dissimilar morph image relative to the other three stimuli (e.g., stimuli A, B, C and F). In the

Sys_trans condition, participants were preexposed to all six morph stimuli in a systematic order (i.e., A, B, C, D, E and F), and in the Scram_trans condition, participants were preexposed to all six morph stimuli in a fixed scrambled order (i.e., A, E, C, D, B and F). Following stimulus preexposure, a generalisation test was given: here, participants were simply asked to rate how likely they thought it was that one of the stimulus endpoints (e.g., stimulus F) shared a particular property of the other stimulus endpoint (e.g., stimulus A).

Graf (2002) found that the amount of morph transformation systematically influenced participants' categorisation performance, such that judging whether two stimuli were from the same category worsened with increasing transformation distance. Given Graf's (2002) finding, the following predictions were made based on the factors outlined above: if transformational knowledge encourages the incidental classification of stimuli into the same category (Zaki & Homa, 1999), then one would expect participants in condition Sys_trans to show an increased level of property generalisation relative to participants in either the Baseline or Surprise conditions. Moreover, as Zaki and Homa state, "if subjects are acquiring transformational knowledge and using this knowledge in a categorisation task, then systematic training should result in superior classification and recognition performance compared with random presentation of the transformational items" (1999, p. 77). Zaki and Homa (1999) confirmed this hypothesis, with participants in their Experiment 1 showing significantly better classification accuracy following systematic, as opposed to scrambled (random), training (Zaki & Homa, 1999). Consequently, one would also expect participants in condition Sys_trans to show an increased level of property generalisation relative to participants in condition Scram_trans. As noted earlier, however, it is also possible that systematic transformational knowledge may lead participants to view the stimuli categorically (due to the introduction of a category boundary at some point along the morph continuum; see Newell & Bülthoff, 2002). If such behaviour occurred, then one would expect the opposite results to those described above: participants in condition Sys_trans should show a reduced level of property generalisation relative to participants in the other conditions.

Assuming a surprise-driven category invention mechanism in spontaneous categorisation (Clapper & Bower, 1994, 2002; Love et al., 2004), one clear prediction is made. That is, given the skewed nature of the stimulus set, participants in the Surprise condition should show a reduced level of property generalisation relative to participants in the other three conditions. If no incidental categorisation occurs, however, and instead the principles of perceptual learning operate (Gibson, 1969), then one would predict that participants in conditions Sys_trans and Scram_trans should show a reduced level of property generalisation relative to participants in the Baseline and Surprise condition. Moreover, one would expect participants in the Surprise condition to show a reduced level of property generalisation relative to participants in the Baseline condition. That is, based on the principles of perceptual learning, the greater the amount of stimulus exposure, the greater the reduction in property generalisation should be.

3.2.2 Method

3.2.2.1 Participants

Sixty-four Cardiff University students took part for course credit. 16 participants were allocated to each of the four preexposure conditions detailed in Table 1.

Table 1. *The four conditions employed to assess the influence of within-category similarity structure on incidental stimulus classification in Experiment 6.*

Condition	Preexposure	Conditioning	Test
Baseline	A / - / - / - / - / F	A+	F
Surprise	A / B / C / - / - / F	A+	F
Sys_trans	A / B / C / D / E / F	A+	F
Scram_trans	A / E / C / D / B / F	A+	F

Note. A, B, C, D, E and F correspond to renderings of the 1%, 20%, 40%, 60%, 80% and 100% images along a morph continuum. + denotes the application of a particular property to a stimulus.

3.2.2.2 Stimuli

The stimuli were individually rendered images taken with permission from Hahn et al. (2009). These stimuli were originally created by morphing together two objects from the same basic level category (Rosch et al., 1976), using 3ds max™

software (Autodesk, Munich, Germany). The basic level objects were taken from five biological categories (bird, fish, head, mushroom, starfish, turnip) and one artefact category (light bulb; see Figure 15). For every category, two objects formed the end points of each morph continuum (the 1% and 100% morph stimuli), from which 20%, 40%, 60% and 80% morph images were rendered. All morph images had a size of 256 × 256 pixels and were presented in greyscale on a 15-in. computer monitor. Participants were seated at approximately arms length from the monitor for the duration of the experiment.

The use of topological (morphing) transformations was chosen as it allows the use of highly realistic experimental materials, and it also affords parametric variation in an object's shape (Hahn et al., 2009).

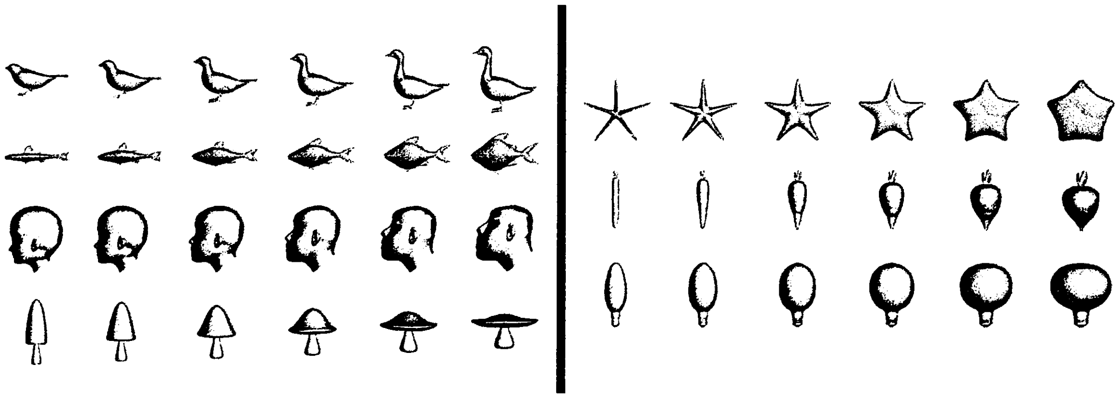

Figure 15. Illustration of the morph stimuli used in the human experiments described in this chapter. The morph continuum was created by morphing between two stimuli from the same basic level category. The stimuli shown here are the 1%, 20%, 40%, 60%, 80%, and 100% morphs, respectively.

3.2.2.3 Design and procedure

A 4 (exposure condition) × 7 (object category) mixed model design was employed. Exposure condition was manipulated as a between-participants factor, and participants in all conditions were exposed to the seven different object categories. On a given trial, participants were sequentially preexposed to a set of morph stimuli from one of the object categories. Within each of the four exposure conditions, half of participants received presentations of the morph stimuli in the order A to F, and half of participants received presentations of the morph stimuli in the order F to A.

Each stimulus was presented for 3000 ms, and the temporal contiguity between the presentation of stimulus A and the presentation of stimulus F was held constant across exposure conditions by introducing a fixation cross when no morph (object) stimulus was scheduled to be presented. Within the subconditions created by the previous counterbalancing operation applied in each exposure condition, following a 1000 ms inter-stimulus interval (blank screen), half of participants were then presented with stimulus A, and half of participants were then presented with stimulus F. Situated above the stimulus was a sentence that informed participants about a particular property that the stimulus had: for example, "This person comes from a small, remote island in the Pacific Ocean". This information remained on the screen until the space bar was pressed, at which point participants were immediately presented with the test screen. On the test screen, participants were simply asked to rate on a scale from 1 (very unlikely) – 9 (very likely) how likely they thought it was that the stimulus now presented to them shared the property of the previously seen stimulus. If participants had previously been presented with stimulus A, then at test, they were presented with stimulus F, and if they had previously been presented with stimulus F, then at test, they were presented with stimulus A. The 1 – 9 rating scale was continuously presented beneath the test stimulus, and responses were made using the 1 – 9 keys on the top of a standard computer keyboard. A 1000 ms inter-trial interval (blank screen) separated participants' likelihood ratings and their preexposure to the next object category. Exposure to the seven object categories was random for all participants in each of the four exposure conditions.

3.2.3 Results

Figure 16 shows the results of the generalisation test: the overall mean likelihood ratings that the test stimulus shared the property of the previously seen stimulus, split by preexposure condition. Inspection of this figure reveals that, overall, participants in the Surprise condition reported lower mean likelihood ratings than participants in the other three preexposure conditions; overall likelihood ratings in the other three conditions were all very similar.

Due to a lack of homogeneity of variances between conditions (Levene's test of homogeneity of variances, $F(3, 60) = 5.23$, $p < .003$), the Brown-Forsythe correction for ANOVA was applied. A one-way ANOVA confirmed that there was an overall effect of preexposure condition, $F(3, 40.51) = 2.85$, $p < .05$, $\eta^2 = .12$.

Dunnett T3 post-hoc tests (equal variances not assumed) revealed that, overall, participants in the Surprise condition reported significantly lower mean likelihood ratings than participants in the Baseline condition ($p < .05$, $r = .35$). No other post-hoc comparisons were significant (all *ps* > .05).

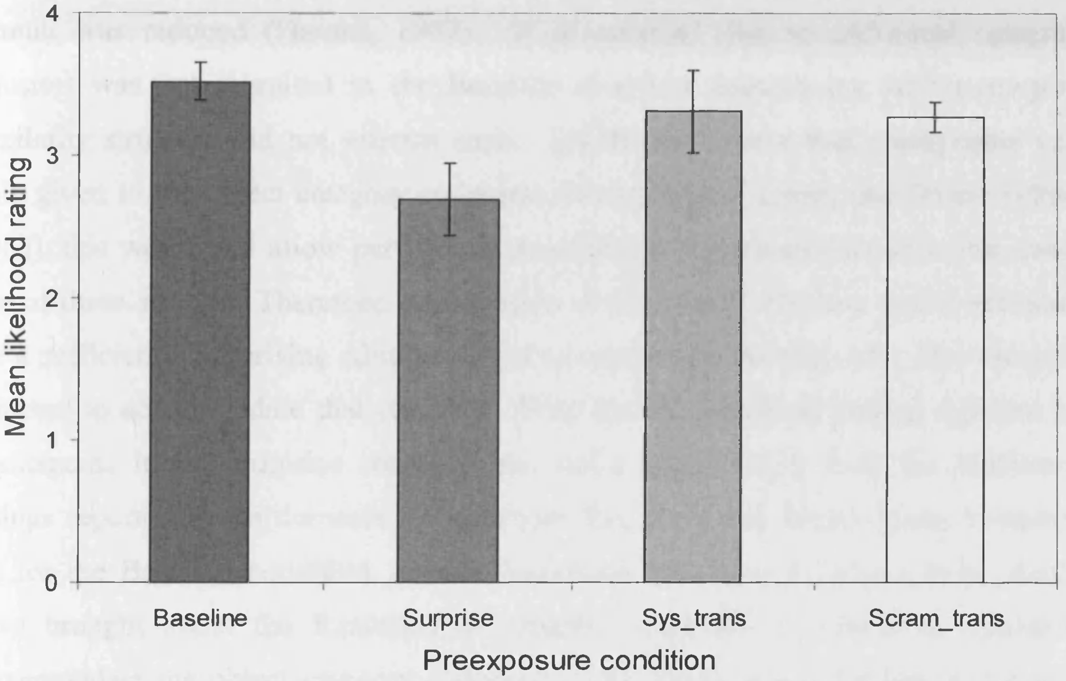

Figure 16. Results of Experiment 6: overall mean likelihood ratings over the seven object categories, plotted by preexposure condition. Error bars indicate the standard error.

3.2.4 Discussion

Participants in the Surprise exposure condition reported significantly lower likelihood ratings over the seven generalisation tests than participants in the Baseline condition. However, likelihood ratings reported by participants in the Surprise condition did not differ significantly from likelihood ratings reported by participants in conditions Sys_trans and Scram_trans. Moreover, likelihood ratings reported by participants in the Baseline condition also did not differ significantly from those reported by participants in conditions Sys_trans and Scram_trans.

The results of Experiment 6, therefore, are broadly consistent with the predictions of a surprise-driven category invention mechanism operating in incidental categorisation (Clapper & Bower, 1994, 2002; Love et al., 2004). This assumes that the within-category similarity structure (i.e., distributional properties) of the Surprise condition encouraged participants to recruit an extra category (cluster) in which to accommodate the lone distinct stimulus. That is, in the Surprise condition, it is assumed that stimulus A was incidentally classified into a different category to that of stimulus F. Consequently, the amount of property generalisation between these stimuli was reduced (Harnad, 1987). It is assumed that an additional category (cluster) was not recruited in the Baseline condition because the within-category similarity structure did not warrant such. Specifically, given that preexposure was only given to the object category endpoints, according to Clapper and Bower (1994, 2002), this would not allow participants to establish a particular set of norms about one of these stimuli. Therefore, presentation of the second stimulus would not make for a sufficiently surprising stimulus event to warrant the creation of a new category (cluster) to accommodate that stimulus. Why did the likelihood ratings reported by participants in the Surprise condition not differ significantly from the likelihood ratings reported by participants in conditions Sys_trans and Scram_trans, however? As for the Baseline condition, neither conditions Sys_trans or Scram_trans should have brought about the formation of separate categories in which to separately accommodate the object category endpoints. One likely reason for this is due to a small influence of perceptual learning operating in condition Sys_trans and Scram_trans, which reduced the perceived similarity between the object category endpoints (i.e., stimuli A and F). Consequently, this reduction in similarity led to a concomitant decrease in the amount of property generalisation between stimuli A and F of each object category (see Pavlov, 1927; Shepard, 1987) in conditions Sys_trans and Scram_trans, relative to the Baseline condition.

Comparing the Baseline condition to condition Sys_trans and Scram_trans, it is apparent that transformational knowledge did not enhance the level of property generalisation between the object category endpoints by increasing the similarity of these stimuli (A and F). Such an increase in stimulus similarity and property generalisation between stimuli A and F was expected based on the assumption that transformational knowledge encourages the perception that two different, but similar stimuli should be 'classified together' (see Zaki & Homa, 1999). Indeed, as noted above, it appears that there was a small influence of perceptual learning in conditions Sys_trans and Scram_trans, leading to a slight numerical reduction in likelihood ratings for these conditions relative to the Baseline condition. Moreover, there is no evidence to suggest that systematic transformational knowledge (condition Sys_trans) influenced participants' response behaviour differently to non-systematic transformation knowledge (condition Scram_trans; cf. Zaki & Homa, 1999). Equally, there is no evidence to suggest that stimuli in condition Sys_trans were perceived categorically (cf. Newell and Bülthoff, 2002). If this condition had encouraged categorical perception (Harnad, 1987), then ratings in this condition should have mirrored those reported in the Surprise condition. It is possible, of course, that in condition Sys_trans, the opposing influences of transformational knowledge and categorical perception may cancelled one another out. This situation would, therefore, have left participants in condition Sys_trans uncertain about the classificatory status of stimuli A and F in each object category, which would have mirrored the uncertainty felt by participants in the Baseline condition and condition Scram_trans.

One problem with concluding that the results of Experiment 6 were driven by a surprise-driven category invention mechanism, operating specifically on within-category similarity structure, is that the Surprise condition also had a distinct temporal structure. That is, while the three stimuli with the highest perceptual similarity were presented in a temporally contiguous manner, a temporal gap of six seconds separated presentation of the distinct stimulus from the highly similar stimuli. It is possible, therefore, that it was this temporal discontinuity, rather than the perceived perceptual discontinuity, that encouraged the formation of a new category (cluster) so as to accommodate the distinct stimulus separately from the highly similar stimuli. This, of course, would have led to the decrease in property generalisation found between stimuli A and F in the Surprise condition, relative to the other three conditions. As documented earlier, a number of authors have found that the temporal dynamics of stimulus preexposure can influence later stimulus generalisation (Bennett & Mackintosh, 1999; see also, Chantrey, 1972, 1974). To assess the influence of this temporal discontinuity in producing the results of Experiment 6, Experiment 7 was undertaken.

3.3 Experiment 7

3.3.1 Introduction

The design of Experiment 7 is summarised in Table 2 below. Participants were allocated to one of two exposure conditions: the first condition was the Baseline condition of Experiment 6. The second condition (Surprise_2) was similar to the Surprise condition of Experiment 6, with the exception that the stimuli were now preexposed in an even temporally spaced manner. That is, a 2000 ms temporal delay separated presentation of each of the four stimuli. Consequently, if the temporal discontinuity contained within the Surprise condition of Experiment 6 was critical in producing the significant difference found in Experiment 6, then the likelihood ratings given in the Surprise_2 condition should not differ significantly from those given in the Baseline condition. If, however, the significant difference found in Experiment 6 was the result of the perceived perceptual discontinuity contained within the Surprise condition, then participants in the Surprise_2 condition should still report significantly lower likelihood ratings than participants in the Baseline condition.

3.3.2 Method

3.3.2.1 Participants

Thirty-two Cardiff University students took part for a small payment of £2. 16 participants were allocated to the Baseline condition and 16 participants were allocated to condition Surprise_2 (see Table 2).

Table 2. *The two conditions employed to assess the influence of within-category similarity structure on incidental stimulus classification in Experiment 7.*

Condition	Preexposure	Conditioning	Test
Baseline	A / - / - / - / F	A+	F
Surprise_2	A / B / C / F	A+	F

Note. A, B, C, D, E and F correspond to renderings of the 1%, 20%, 40%, 60%, 80% and 100% images along a morph continuum. + denotes the application of a particular property to a stimulus.

3.3.2.2 Stimuli, design and procedure

The stimuli employed were those used in Experiment 6. The design and procedure was also that used for Experiment 6, with the following exception: during the preexposure phase of the Surprise_2 condition, presentations of the morph stimuli were separated by a 2000 ms long fixation cross. This not only eliminated the temporal discontinuity present in the Surprise condition of Experiment 6, but it also maintained an equivalent temporal spacing between presentations of the object category endpoints across the two conditions.

3.3.3 Results

Figure 17 shows the results of interest: the overall mean likelihood ratings split by preexposure condition. Inspection of Figure 17 shows that, overall, participants in the Surprise_2 condition reported lower likelihood ratings than participants in the Baseline condition. Due to a violation of normality in the Baseline condition (Shapiro-Wilk test of normality, $p < .007$), the nonparametric Mann-Whitney U test was conducted on the data. This test revealed that overall likelihood ratings given in the Surprise_2 condition were significantly lower than those given in the Baseline condition, $U(16, 16) = 56.50, p < .008, r = .49$[10].

[10] ANOVA also confirmed this difference to be significant, $F(1, 30) = 6.14, p < .02$.

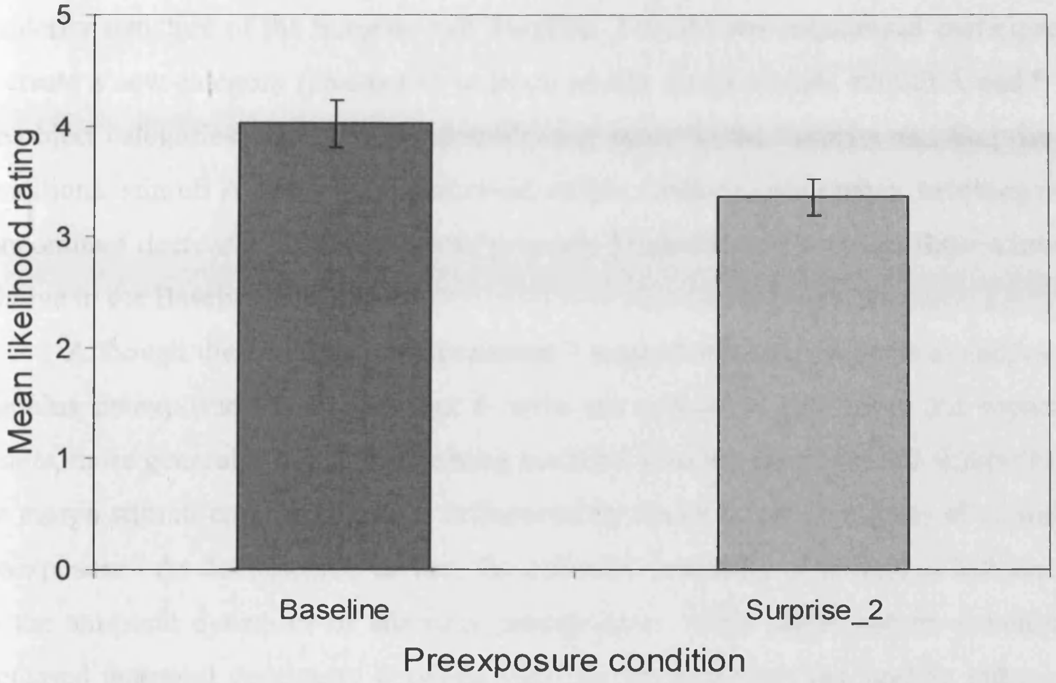

Figure 17. Results of Experiment 7: overall mean likelihood ratings over the seven object categories, plotted by preexposure condition. For purposes of consistency, mean ratings rather than median ratings are presented. Error bars indicate the standard error.

3.3.4 Discussion

The results of Experiment 7 confirm those of Experiment 6. That is, participants in condition Surprise_2 reported lower likelihood ratings over the seven object categories relative to participants in the Baseline condition. Consequently, the current results provide no evidence to support the claim that it was the temporal discontinuity present during stimulus preexposure in the Surprise condition of Experiment 6 that was influential in producing the significant difference between the Baseline and Surprise conditions of that experiment. Indeed, the effect size observed in the present experiment was actually numerically greater than that observed for the significant post hoc contrast between the Baseline condition and the Surprise condition in Experiment 6.

Taken collectively, I would argue that the results of Experiments 6 and 7 support the predictions of a surprise-driven category invention mechanism in incidental categorisation, which operates on within-category similarity structure

(Clapper & Bower, 1994, 2002; Love et al., 2004). That is, only the within-category similarity structure of the Surprise and Surprise_2 conditions encouraged participants to create a new category (cluster) so as to separately accommodate stimuli A and F of the object categories. Due to this 'classification apart' in the Surprise and Surprise_2 conditions, stimuli A and F were perceived as less similar to each other, resulting in a concomitant decrease in the amount of property generalisation between these stimuli, relative to the Baseline condition.

Although the findings of Experiment 7 suggest that the temporal dynamics of stimulus preexposure in Experiment 6 were not critical in producing the reported results, more generally it is an interesting question whether the perceived similarity of the morph stimuli employed here is influenced by the temporal contiguity of stimulus preexposure. As documented earlier, the effective similarity of stimuli is influenced by the temporal dynamics of stimulus preexposure: while under certain conditions increased temporal contiguity between stimulus presentations can lead to enhanced perceptual learning – as in the case of the intermixed versus blocked effect, for example – under different conditions, increased temporal contiguity between stimulus presentations can encourage sensory preconditioning (see Bennett & Mackintosh, 1999; Honey & Bateson, 1996). Specifically, Hall (1991, p. 235; see also McLaren & Mackintosh, 2000) has argued that sequential stimulus exposure that occurs at very high temporal contiguity should be most likely to lead to stimuli becoming associatively linked (or 'classified together'; see Bateson & Chantrey, 1972). As a consequence of this, the effective similarity of stimuli should increase when stimuli are sequentially presented at high temporal contiguity, relative to some baseline. The majority of evidence for this, however, is from studies that have employed rather simple stimuli. Therefore, I was keen to assess whether the temporal contiguity of stimulus exposure would influence the perceived similarity of the complex, naturalistic stimuli employed in Experiments 6 and 7 of this chapter. In the following section (Section 3.4.1) I explore more fully why temporal contiguity should, under certain conditions, increase stimulus similarity.

3.4 Experiment 8

3.4.1 Introduction

Why should temporally contiguous, sequential stimulus exposure increase the perceived similarity between the object category endpoints (A and F) of the previously employed stimuli? One possibility is that presenting stimulus F immediately after stimulus A, for example, will permit the formation of an association between these stimuli. This A – F association means that the presentation of stimulus A will evoke a representation of stimulus F, causing these stimuli to be perceived equivalently, and therefore increasing their effective similarity (Hall, 1991; see also, Honey & Bateson, 1996). This position is similar in kind to the proposal of Bateson and Chantrey (1972; see also Chantrey, 1974), in which they suggested that two stimuli presented in close temporal contiguity in the same context will be 'classified together'. In contrast, stimuli that are not presented in close temporal contiguity will be 'classified apart'. One way of conceptualising this 'classification together' is in terms of the formation of a blended representation of stimulus A and stimulus F. That is, when stimulus A and stimulus F are presented in close temporal contiguity, this may establish the representation AF (Hall, 1991; see also, Pearce, 1987). Consequently, whenever stimulus A and stimulus F are attended to, these stimuli will evoke the blended representation AF, increasing their perceived similarity to each other.

Interestingly, the predictions made above do not follow from the theorising of Gibson (1969). Rather, she suggested that increasing the temporal contiguity of stimulus preexposure should result in stimuli becoming less similar. This is because being able to compare stimuli closer together in time will be particularly effective in encouraging a process of stimulus differentiation, in which attention is drawn to the unique features of the stimuli and away from their common features. Indeed, Gibson and Walk (1956) found that prolonged simultaneous stimulus exposure does result in better discrimination learning at a later time than no stimulus exposure. Recently, Mundy et al. (2007, 2009) have shown that simultaneous stimulus preexposure does lead to better discrimination learning than sequential stimulus preexposure. However, this facilitation effect from simultaneous preexposure appears to be sensitive to the similarity of the stimuli being exposed. That is, while highly similar stimuli have been shown to become more discriminable following simultaneous preexposure, Mundy et al. (2007) found no such effect for highly discriminable stimuli. This result is consistent with the idea that stimulus exposure engages both comparator and associative processes, and that under certain circumstances, it is simply a question of which, if any, process wins out (see Honey & Bateson, 1996; Honey et al., 1994).

When considering the morph stimuli used in the previous experiments, it therefore seems reasonable to suppose the following: given that stimuli A and F of each object category are readily discriminable from the outset, temporally contiguous exposure to just these stimuli should result in the formation of an excitatory association between them; in other words, they should become 'classified together'. In contrast, when a temporal delay is introduced between exposure to stimuli A and F, the formation of an excitatory association between these stimuli is far less likely (or, at least, any excitatory association will form more weakly). Furthermore, when stimulus exposure is temporally contiguous overall, but there is a delay between exposure to stimuli A and F, due to the introduction of transformational knowledge, both associative and comparative processes will likely be active. That is, while excitatory associations should form between each contiguously presented pair of stimuli (e.g., between stimulus A and stimulus B, and, between stimulus B and stimulus C, and so on...), the fact that each presented stimulus is not readily discriminable from its neighbours means that a comparison process, or stimulus differentiation, should also be encouraged. If the influence of these two processes is relatively balanced, then the similarity of stimuli A and F may remain relatively unchanged. The upshot of all this is that one would predict that the perceived similarity of stimuli A and F should be greater following highly contiguous exposure to these two stimuli, relative to stimulus exposure that incorporates a delay between the presentation of stimulus A and stimulus F. Of particular interest is the comparison between highly contiguous stimulus exposure and stimulus exposure that incorporates transformational knowledge. If the prior prediction were supported, this would be contradictory to the proposal of Zaki and Homa (1999). To recapitulate, they suggested that transformational knowledge should encourage the classification of two distinct stimuli into the same category, which should increase stimulus similarity (Harnad, 1987).

Experiment 8, therefore, sought to assess whether the three conditions of stimulus exposure outlined in the previous paragraph differentially affected the perceived similarity between the previously employed object category endpoints (i.e., stimuli A and F). To this end, participants were allocated to one of three preexposure conditions, detailed in Table 3. Specifically, participants in the Baseline condition received preexposure to the stimuli in the same manner as participants in the Baseline condition of Experiments 6 and 7. Participants in condition Sys_trans received stimulus preexposure in the same manner as participants in condition Sys_trans of Experiment 6. In the newly introduced Contiguous condition, participants received preexposure only to stimuli A and F of the object categories (as for participants in the Baseline condition), but in this condition, presentation of the second stimulus (e.g., F) followed immediately after presentation of the first stimulus (e.g., A). Following stimulus preexposure, participants were simply asked to rate how similar stimulus F was to stimulus A (or vice versa) on scale from 1 (very dissimilar) to 9 (very similar).

3.4.2 Method

3.4.2.1 Participants

Forty-eight Cardiff University students took part either for course credit or a small payment of £2. 16 participants were allocated to each condition (see Table 3).

Table 3. *The three conditions employed to assess incidental stimulus classification in Experiment 8.*

Condition	Preexposure	Test
Baseline	A / - / - / - / - / F	A-F
Sys_trans	A / B / C / D / E / F	A-F
Contiguous	A / F	A-F

Note. A, B, C, D, E and F correspond to renderings of the 1%, 20%, 40%, 60%, 80% and 100% images along a morph continuum. + denotes the application of a particular property to a stimulus.

3.4.2.2 Stimuli, design and procedure

The same stimuli and design employed in Experiments 6 and 7 were used. Participants in the Baseline and Sys_trans conditions received stimulus preexposure that was identical to the Baseline and Sys_trans conditions in Experiment 6.

Participants in the Contiguous condition received preexposure to the object category endpoints (stimuli A and F) in a temporally contiguous fashion. That is, presentation of the second stimulus (e.g., F) followed immediately after presentation of the first stimulus (e.g., A). As for Experiments 6 and 7, all stimuli were presented for 3000 ms during preexposure. In each of the three exposure conditions, half of participants received stimulus preexposure in the order A to F, and half of participants received stimulus preexposure in the order F to A. Following stimulus preexposure, a 1000 ms inter-stimulus interval (blank screen) separated presentation of the test screen, on which was presented stimulus A and stimulus F. Within the subconditions created in each exposure condition following the previous counterbalancing operation, half of participants saw stimulus A surrounded by a red border on the test screen, and half of participants saw stimulus F surrounded by a red border on the test screen. Within each of the subconditions created by the previous counterbalancing operations, half of participants received presentations of stimulus A on the left-hand side of the test screen and presentations of stimulus F on the right-hand side of the test screen, and half of participants received the reverse. On the test screen, participants were simply asked to rate how similar they thought the object framed in red was to the object not framed in red, using a 1 (very dissimilar) to 9 (very similar) rating scale presented at the bottom of the test screen. Participant responses were made using the keys "1" through "9" on the top of a standard keyboard. Following a response, a 1000 ms inter-trial interval (blank screen) separated participants' exposure to the next object category. Exposure to the seven object categories was again random for all participants in each of the three exposure conditions employed here.

3.4.3 Results

Figure 18 displays the results of interest: participants' overall mean similarity rating over the seven object categories, split by preexposure condition. As predicted, overall similarity ratings were higher in the Contiguous condition than in the Baseline and Sys_trans conditions. Overall similarity ratings in the Baseline condition differed little from those in the Sys_trans condition. A one-way ANOVA revealed that there was a significant effect of exposure condition, $F(2, 45) = 7.31, p < .003$. Tukey HSD post-hoc tests revealed that, overall, participants in the Contiguous condition reported significantly higher ratings of similarity than participants in the Baseline condition (*p*

< .05) and Sys_trans condition (p < .002). Overall similarity ratings did not differ significantly between the Baseline and Sys_trans conditions (p > .05).

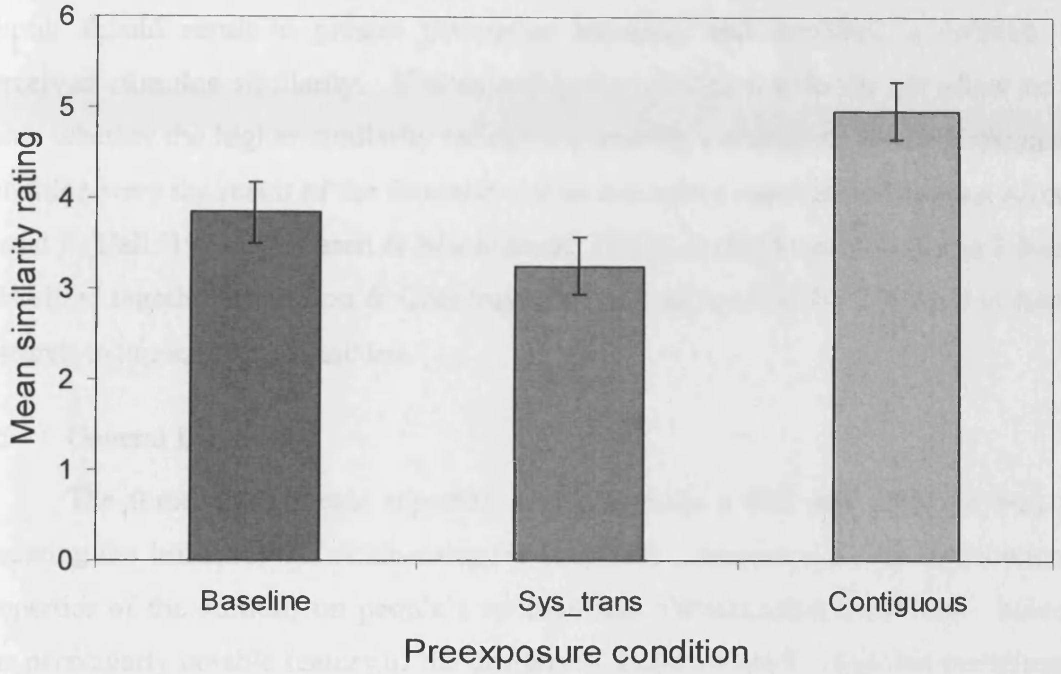

Figure 18. Results of Experiment 8: overall mean similarity ratings over the seven object categories, plotted by preexposure condition. Error bars indicate the standard error.

3.4.4 Discussion

In agreement with the proposal of Hall (1991, p. 235), the perceived similarity of stimuli A and F was rated highest in the Contiguous exposure condition. Of particular interest is the finding that ratings of similarity in the Contiguous condition were significantly higher than those in the Sys_trans condition. With respect to this contrast at least, the presumed transformational knowledge in condition Sys_trans actually had a negative influence on the perceived similarity of stimuli A and F. As in Experiment 6, the Sys_trans condition did not differ from the Baseline condition, demonstrating that under conditions of brief stimulus exposure, transformational knowledge does nothing to increase the perceived similarity of two different, but similar stimuli. To reiterate, these latter findings contrast with the arguments of Zaki and Homa (1999), who proposed that transformational knowledge should encourage two different, but similar stimuli to be classified into the same category; a process that should increase perceived stimulus similarity (Harnad, 1987). The results of Experiment 8 also do not support the predictions of Gibson (1969). To recapitulate, she argued that increasing the temporal contiguity between preexposure to different stimuli should result in greater perceptual learning, and therefore, a decrease in perceived stimulus similarity. Unfortunately the present results do not allow us to know whether the higher similarity ratings reported by participants in the Contiguous condition were the result of the formation of an excitatory association between stimuli A and F (Hall, 1991; McLaren & Mackintosh, 2000), or due to stimuli A and F being 'classified together' (Bateson & Chantrey, 1972; Chantrey, 1974). I leave it to future research to unpick this distinction.

3.5 General Discussion

The three experiments reported above provide a fast and effective way of assessing the influence of within-category similarity structure (i.e., the distributional properties of the stimuli) on people's spontaneous classification behaviour. Indeed, one particularly notable feature of the designs of Experiments 6 – 8 is that participants only received a single presentation of each scheduled stimulus during preexposure. Two main findings were evident from Experiments 6 and 7: First, transformational knowledge did not increase the amount of property generalisation between the endpoints of the object categories (i.e., A and F). Second, when there was structural discontinuity that could be perceived within the preexposed stimulus set, this surprising event led to a reduction in the amount of property generalisation between stimuli A and F. The latter result supports the assumption that a surprise-driven category invention mechanism operates within human spontaneous categorisation, and that participants in the Surprise condition likely came to classify stimuli A and F into different categories/ clusters (see Clapper & Bower, 1994, 2002; Love et al., 2004; also, Gureckis & Goldstone, 2008). Such 'classification apart' would have resulted in a decrease in the perceived similarity of stimuli A and F, and as such, a concomitant reduction in the amount of generalisation between these stimuli (Harnad, 1987). These findings support work by Grand, Close, Hale and Honey (2007), which has shown that stimulus similarity commands an important influence over the amount of associative transfer between two stimuli. For example, in one experiment, Grand et al. (2007) showed that one observes better associative transfer of a conditioned response between two stimuli when the stimuli are similar (e.g., AX and BX), compared to when they are dissimilar (e.g., CX and DY). Importantly, the pattern of results of Experiment 6 shows that stimulus generalisation was not simply determined by the amount of stimulus exposure. Experiment 8 compared the rated similarity of stimuli A and F following either preexposure only to stimuli A and F or to stimuli A, B, C, D, E and F as a systematic transformation. When the temporal delay between presentation of stimuli A and F was equated between conditions, no difference in rated similarity was observed. However, when stimuli A and F were presented contiguously, participants rated the similarity of these stimuli more highly than participants that received preexposure to stimuli A and F in a non-contiguous fashion, regardless of whether transformational knowledge was present or not.

In conclusion, Experiments 6 and 7 build on the work of Chapter 2 in further demonstrating that, in humans, the similarity structure of a set of stimuli influences their spontaneous classification. That is, perceived discontinuities in the environment appear to help guide people's identification of category structure (Anderson, 1991; Malt, 1995; Rosch & Mervis, 1975); of course, other factors are clearly influential in determining this process too (e.g., temporal contiguity, see Experiment 8, and general knowledge, see Heit, 1997; Murphy, 2002). This conclusion is mirrored in a cross-cultural review of natural and artefactual categorisation carried out by Malt, in which she concluded that "there is structure in the environment that is perceived in a universal fashion by human categorizers" (1995, p. 128). While Malt (1995) specifically makes her conclusions with respect to human categorisation, the term universal is an evocative one: if structure in the environment is indeed perceived in a consistent manner across many different cultures by humans, then it seems plausible to suppose that other, nonhuman animal species may also be sensitive to this same structure. It is possible, therefore, that any tendency for spontaneous categorisation in nonhuman animals may be driven by similar principles as for human spontaneous categorisation. Malt (1995) notes, of course, that human categorisation must result from an interaction between the environment and the classifier, concluding that structure alone is not sufficient to determine categorisation. Consequently, it is equally plausible that any tendency for spontaneous categorisation in nonhuman animals may be qualitatively different from that of human spontaneous categorisation.

One of the particularly nice features about the experimental design of Experiments 6 and 7 is that it can be readily transposed and applied to an assessment of incidental categorisation in nonhuman animals. The assessment of incidental categorisation in nonhuman animals is, I believe, a particularly interesting question. First, it is still an open question whether nonhuman animals engage in any meaningful form of spontaneous categorisation at all. Second, if nonhuman animals do engage in incidental categorisation, is this determined by the same mechanisms that guide human incidental categorisation (i.e., a surprise-driven category invention mechanism)? The incidental categorisation procedure introduced in Experiment 6 provides one possible direct test of these important questions. If nonhuman animals do engage in incidental categorisation in a manner that is consistent with the human results of Experiment 6, then this would suggest a common ancestry in the development of our spontaneous categorisation abilities, determined by the perceived structural properties within the environment. Consequently, the role of the classifier is somewhat downplayed. If, however, no evidence for incidental categorisation is found in the incidental classification task, then this would suggest a more primary role for the classifier. In the remaining half of this chapter, therefore, I sought to investigate incidental categorisation in one nonhuman animal species; namely, the rat. To reiterate, the aims of these experiments were to i) assess whether rats engage in incidental (spontaneous) categorisation, and ii) if they do, to determine whether this incidental categorisation occurs in a manner that is consistent with the human results of Experiment 6 (i.e., guided by a surprise-driven category invention mechanism).

Before presenting these experiments in rats, however, I will first briefly review the state of nonhuman animal categorisation research. Moreover, I will also review some of the factors that may influence whether nonhuman animals come to 'classify together' or 'classify apart' a set of stimuli. As will be shown, these factors are not dissimilar to the factors that may have influenced incidental categorisation in humans, which were identified earlier.

3.6 Categorisation in nonhuman animals

In his review of nonhuman categorisation, Herrnstein (1990, p. 138) states that categorisation has "turned up at every level of the animal kingdom where it has been competently sought". Indeed, under supervised task conditions, nonhuman animals are able to learn complex discriminations that resemble human categorisation. For example, Herrnstein et al. (1976) showed that pigeons were able to learn a complex discrimination between different scenes presented on photographic slides; half of these scenes contained pictures of trees (which were not particularly prominent), and half did not. Slides were presented sequentially, and only those scenes that contained pictures of trees were rewarded with food following a peck at a response key. More than 500 different scenes made up the pool from which the slides could be selected. While the large number of different slides used makes it unlikely that pigeons simply remembered each slide and its associated outcome separately (but, see Vaughan & Greene, 1984), this possibility was ruled out in a further test of generalisation. Specifically, pigeons were shown to respond correctly to novel photographic slides containing either trees or no trees.

Similarly, Cerella (1979) showed that pigeons could learn a complex discrimination between different leaf types. Using 80 slides of different kinds of leaves, pigeons came to correctly respond in the presence of silhouetted oak leaves, but not in the presence of silhouetted non-oak leaves. Critically, this behaviour was also shown to generalise to novel silhouetted oak leaves. This proficiency for learning complex discriminations has been documented in many different animal species using both natural and artificial stimuli (e.g., Marsh & MacDonald, 2008; Mercado III, Orduña, & Nowak, 2005; Morgan, Fitch, Holman, & Lea, 1976; Schrier, Angarella, & Povar, 1984; Schrier & Brady, 1987; Vogels, 1999; Vonk & MacDonald, 2002, 2004). Porter and Neuringer (1984) have further shown that pigeon's highly adept ability to learn certain discriminations is not confined to the visual domain. They showed that pigeons were able to learn an auditory discrimination between compositions by J.S. Bach and Stravinsky. There is also some evidence for discrimination learning based on the more abstract, relational notion of 'sameness': following discrimination learning between a set of stimuli based on the human concept 'same-different', pigeons, corvids, rhesus monkeys, baboons and chimpanzees have all shown successful transfer of same-different learning to novel stimuli (e.g. Young & Wasserman, 1997; Wilson, Mackintosh, & Boakes, 1985; Mishkin, Prockop, & Rosvold, 1962; Fagot, Wasserman, & Young, 2001; Oden, Thompson, & Premack, 1988; respectively). However, these studies have also shown that the manner in which nonhuman animals respond to same-different discriminations is somewhat different to the way humans respond to same-different discriminations. For example, whereas pigeons have been found to respond in a manner that is continuous in nature (e.g., Young & Wasserman, 1997; but see Premack, 1983; Thompson, Oden, & Boysen, 1997) – that is, they respond proportionally to intermediate arrays that contain some same and some different items – humans have been found to respond categorically; responding, for the most part, "different" to the intermediate arrays (Young & Wasserman, 2001).

In summary, the above research has provided compelling evidence that nonhuman animals can come to treat an experimentally defined set of stimuli in the same way. However, it remains unclear from these studies whether nonhuman animals engage in meaningful stimulus grouping (i.e., categorization). Rather, the fact that nonhuman animals come to treat a set of stimuli in the same way may simply reflect the fact that each of the stimuli, which form an experimentally derived category, has been individually associated with the same outcome. Indeed, Chater and Heyes argue that "Experiments suggesting that animals can form 'equivalence classes' may be mistakenly interpreted as evidence that, contrary to the predictions of standard stimulus generalisation models, animals' stored representations of category members *do* have something in common beyond the fact that each is independently associated with a common response or trial outcome" (1994, p. 216). However, not all experiments appear mistaken in this regard.

Honey and Watt (1998, 1999) had rats engage in a biconditional discrimination task involving two cues (X and Y) and four contexts (A, B, C and D). Presentations of AX and BX (but not CX and DX) signalled the delivery of food, and presentation of CY and DY (but not AY and BY) signalled the delivery of food. Critically, this arrangement meant that each context was paired equally often with the two cues, and with food and the absence of food (that is, everything was equal). Following discrimination training, context A was paired with shock, and context C was not. Honey and Watt (1998, 1999) found that rats showed greater generalisation of the fear response (produced by the shock) to context B than to context D. That is, discrimination training altered the effective similarity of the context stimuli, such that contexts A and B were perceived as more similar than contexts A and D, and contexts C and D were perceived as more similar than contexts B and C. This result is not predicted by prominent configural accounts of learning (e.g., Pearce, 1987, 1994), which would assume that contexts A, B, C and D would share an equivalent amount of similarity following the given discrimination training. One interpretation of these findings has therefore assumed a form of stimulus grouping: when similar compounds (e.g., AX and BX) are followed by the same outcome (e.g., food), their components (i.e., *a*, *b* and *x*) come to address a shared configural unit within a connectionist network, and when similar compounds (e.g., AX and DX) are followed by different outcomes, their components (i.e., *a*, *d* and *x*) come to address different configural units (see Allman, Ward-Robinson, & Honey, 2004; Honey & Ward-Robinson, 2002). As in humans (see Harnad, 1987), stimulus grouping appears to alter the effective similarity of stimuli in nonhuman animals, such that stimuli that have been grouped together become more similar, and stimuli that have been grouped apart become less similar. Honey and Watt's (1998, 1999) results are important, therefore, as this change in the effective similarity of stimuli following stimulus grouping (classification) formed the basic premise for Experiments 6 and 7.

So, supervised training can elicit both complex discriminatory behaviour and stimulus grouping in nonhuman animals. However, one rather important question still remains: do nonhuman animals engage in such stimulus grouping spontaneously? That is, does a pigeon really spontaneously group together different kinds of trees into some unitary category that is distinct from other environmental stimuli? While it is clear that animals are sensitive to the statistical properties of a supervised learning task, are they implicitly sensitive to these regularities and the structure of their environment? Of course, we know already that mere exposure to stimuli can influence the way in which both humans and nonhuman animals later perceive these stimuli (Hall, 1991). However, perceptual learning experiments have typically not been focused towards an understanding of nonhuman animals' implicit sensitivity to category structure (i.e., the distributional properties of the stimuli that they are exposed to). Consequently, this has limited our understanding of spontaneous categorisation in nonhuman animals. There are, however, a number of studies that at least suggest that nonhuman animals are sensitive to the distributional properties of a set of stimuli, and to stimulus structure.

3.6.1 Transformational information in chicks

A number of situations exist throughout the animal kingdom that may require similar stimuli to be spontaneously classified into the same category. For example, a young chick will obviously view its mother from many different viewpoints and through seasonal changes, and yet it must be able to appreciate that all these snapshot images signify its mother. The mother hen faces an even more challenging problem; Ryan and Lea (1990) have shown that mother hens recognise their offspring, and they must continue to do so throughout the chick's marked maturational development. Of course, family unity can be maintained in a number of ways, which most likely include using auditory and olfactory references (although vocal changes are also associated with chick maturation; Ryan & Lea, 1990). However, Ryan (1982) has shown that chickens are readily able to learn a purely visual discrimination between two object birds, and that this discrimination learning is unaffected by the age of the object birds. She further found that, after being trained to discriminate between the two object birds at a specific age (e.g., 2 days old), chickens showed some ability to generalise their learning to a discrimination that involved the same object birds but at a different age (e.g., up to 43 days old). Based on a series of experiments, and given the marked maturational changes development of the chick, Ryan and Lea concluded that "generalization per se would not be sufficient to allow continued recognition from hatching to independence" (1990, p. 98). What is required, they argue, is a process of *representational updating*, which would compliment the principles of stimulus generalisation.

The process of representational updating could proceed in one of two ways: it could be that the mother hen engages in a process whereby she periodically updates her old representations of each of her offspring by simply overwriting these with new ones. Equally, it is possible that the mother hen might incorporate the maturational changes that her offspring go through into her original representations of each chick. That is, rather than simply replacing a sibling's representation with a more up-to-date version, the mother hen may gradually enlarge her category of "offspring". Ryan and Lea (1990) sought to test between these accounts by using an imprinting procedure in chicks, rather than focusing on the mother hen. Imprinting refers to the phenomenon that chicks are predisposed to move towards and follow the first salient object that they view following birth (normally their mother). In Ryan and Lea's (1990) experiment, 121 chicks were initially imprinted on a string of four table tennis balls (e.g., AAAA), and were then housed in visual isolation, away from other chicks. Chicks in the experimental condition were then subjected to a gradual change in the colour of the string of balls (either from white to brown or vice versa) by replacing the four balls, one at a time, for balls of the opposite colour. One ball was replaced for the opposite colour every four days. Three control groups were used for comparison. The first group consisted of birds that were simply exposed to the imprinted string of balls (i.e., AAAA) over the whole 21 days of the experiment. The second group of birds experienced an abrupt change in stimulus form; that is, chicks were exposed to the imprinted string of balls (i.e., AAAA) for the first 17 days, and then they were exposed to a string of balls of the opposite colour (e.g., BBBB) for the remaining 4 days of the experiment. The third group of birds were only exposed to the imprinted string of balls (i.e., AAAA) for the first four days of the experiment, at which point the balls were removed and not replaced. Subsequently, a preference test was undertaken using a Y-maze: chicks could choose to move towards either the originally imprinted coloured string of balls (i.e., AAAA), or towards the string of balls of the opposite colour (i.e., BBBB). Ryan and Lea (1990) found that while birds in the control groups showed a preference to move towards the originally imprinted string of balls (i.e., AAAA), those birds in the experimental group did not show any choice preference. Of particular interest is the comparison between the experimental group and the control birds that experienced the abrupt change of ball colour at 17 days. As Ryan and Lea (1990) state, if chicks engaged in complete, successive replacement of old representations with new ones, then the chicks that experienced this abrupt change should have shown a preference for the coloured string of balls presented over the final four days, and not for the originally imprinted string of balls. To reiterate, however, chicks that experienced the abrupt change in stimulus form showed a choice preference for the originally imprinted string of balls (i.e., AAAA).

These results, therefore, provide some evidence to support the claim that chicks engage in a process of representational updating, which is afforded by the introduction of transformational information. That is, the intermediate, transformational steps present only in the experimental condition encouraged chicks to expand their initial representation of the imprinted stimulus to include all subsequent changes in its form; or to put this another way, chicks in the experimental condition appear to have spontaneously 'classified together' the various stimulus forms into the same category. It seems, therefore, that transformational information can influence both human (e.g., Zaki & Homa, 1999) and nonhuman animal classification behaviour. This experiment also highlights the importance of stimulus similarity in the process of representational updating. Based on chicks' behaviour in the abrupt change condition of Ryan and Lea's (1990) study, it is apparent that a novel stimulus will only be incorporated into the originally imprinted stimulus representation if it shares some degree of similarity (here in terms of feature overlap) with the imprinted stimulus.

3.6.2 Abstract similarity structure

Other research that has assessed whether nonhuman animals engage in spontaneous categorisation has been conducted in nonhuman primates. One notable reason for this is because other primate species are able to manipulate objects in the same manner as human children. Therefore, the pre-linguistic tasks that have been developed to investigate spontaneous categorisation in infants (most commonly the sequential touching procedure) can be readily translated for use with nonhuman primates. In a review of spontaneous categorisation in chimpanzees and monkeys, Spinozzi has concluded that "the ability to classify objects according to perceptual rules of similarity occurs spontaneously in language-trained chimpanzees" (1996, p. 21). For example, Premack (1976) report that with little or no guidance, two language-trained chimpanzees spontaneously classified novel objects into two containers based either on the object's form or colour: when presented with red and yellow triangles and squares, for example, one chimpanzee was found to shift between an initial partitioning on colour, to a partitioning on form, and so on. Intriguingly, the second chimpanzee partitioned these stimuli only on form (i.e., grouping triangles together and squares together). Other research in non-language trained chimpanzees and monkeys, which has mostly employed the sequential touching procedure, has found that chimpanzees and capuchin monkeys (and to a lesser degree macaque monkeys) manipulate objects drawn from two different classes in a consistent, sequential manner. For example, when given yellow, blue and red sticks, and yellow, blue and red rings, these animals have been found to manipulate and group together all the stick objects first, followed by all the ring objects second (Spinozzi, 1993; Spinozzi & Natale, 1989; Spinozzi et al., 1999). What these studies suggest, therefore, is that chimpanzees and some monkey species appear to spontaneously appreciate the similarity structure of their environmental stimuli (i.e., the similarities and differences among the stimuli), and that they use this structure to guide their behaviour towards these stimuli.

3.6.3 Conclusions

In summary, nonhuman animals, like humans, can engage in complex discrimination behaviour when guided by supervised training. What is more, this supervised training can bring about a change in the effective similarity of stimuli, such that stimuli that are paired with the same outcome come to acquire equivalence, and stimuli that are paired with different outcomes come to acquire distinctiveness (Honey & Watt, 1998, 1999; see also, Honey & Hall, 1989). One interpretation of this change in stimulus similarity assumes that it is based on a form of stimulus grouping, or classification (see Honey & Ward-Robinson, 2002). While it is clear that mere exposure can alter stimulus similarity (Hall, 1991; McLaren & Mackintosh, 2000), and specifically that temporally contiguous preexposure to two stimuli can increase stimulus similarity (e.g., Bateson & Chantrey, 1972), it is not clear, to my mind anyway, that nonhuman animals really do engage in meaningful spontaneous categorisation. Furthermore, whether or not nonhuman animals are sensitive to the category structure contained within a set of stimuli is still a somewhat open question. Yes, the work by Ryan and Lea (1990) and, for example, Spinozzi (see Spinozzi, 1996) is suggestive of the fact that nonhuman animals, like humans, are sensitive to stimulus structure. What is more, this work suggests that this structure is likely to be influential in guiding any spontaneous classification behaviour. However, further research on spontaneous categorisation in nonhuman animals is clearly required in order to establish whether it obeys the same principles as for humans.

Investigating spontaneous categorisation in nonhuman animals is difficult, as categorisation can often only be indirectly confirmed from measuring stimulus generalisation and discrimination performance. However, the experimental design described in Experiment 6 can be readily applied to the study of incidental categorisation in many different animal species. In the second half of this chapter, therefore, I sought to assess incidental categorisation in the rat using the experimental design (and the associated assumptions) detailed in Experiment 6. In particular, I was interested in contrasting the results from rats with the results from humans. So, do rats, like humans, utilise a 'surprise-driven' category invention mechanism in incidental categorisation, or will other factors, such as transformational information and perceptual learning, dominate?

3.7 Experiment 9

3.7.1 Introduction

To investigate the questions posed above, rats were allocated to one of four exposure conditions, where they received preexposure over a number of days to a set of four tone stimuli of 1 kHz, 2 kHz, 3 kHz and 4 kHz (henceforth labelled A, B, C and D, respectively). Specifically, rats allocated to the Baseline condition were preexposed only to the endpoints of the tone continuum (i.e., to stimulus A and stimulus D). Rats allocated to the Surprise condition were preexposed to the endpoints of the tone continuum, plus one of the intermediate stimuli (i.e., either stimuli A, B and D, or, stimuli A, C and D). Rats allocated to the final two conditions were preexposed to all four tone stimuli. However, while rats in condition Sys_trans were preexposed to the four stimuli in a systematic order (e.g., A, B, C and D), rats in condition Scram_trans were preexposed to the four stimuli in a fixed scrambled order (e.g., A, C, B and D). Following stimulus preexposure, an appetitive response was conditioned either to stimulus A or stimulus D. Subsequently, rats received a generalisation test. If rats had received appetitive conditioning to stimulus A, then at test, rats received test presentations of stimulus D, and vice versa. To assess the extent of generalisation, the number of magazine entries made during presentations of the test stimulus was recorded.

As for Experiments 6 and 7, it was assumed that, relative to a baseline, the spontaneous classification of stimuli into the same category would increase later generalisation between the test stimuli, whereas the spontaneous classification of stimuli into different categories would decrease later generalisation between the test stimuli. If rats are sensitive to transformational information through a process of representational updating, then it was expected that rats in condition Sys_trans should show an increased level of generalisation between stimulus A and stimulus D, and therefore show a greater level of responding to the test stimulus, relative to rats in the other three conditions. In contrast, if rats came to perceive the four tone stimuli as members of different categories (through the process of categorical perception), then one would expect rats in condition Sys_trans to show a decreased level of generalisation between stimulus A and stimulus D, and therefore show a reduced level of responding to the test stimulus, relative to rats in the other three conditions. If preexposure to the tone stimuli resulted in perceptual learning, then it was expected that rats in both condition Sys_trans and Scram_trans would show a reduced level of generalisation between stimulus A and stimulus D compared to rats in the Baseline condition and the Surprise condition. However, if rats are sensitive to abstract stimulus structure in the way that humans are, then based on the predictions of a surprise-driven category invention mechanism (Clapper & Bower, 1994, 2002; Love et al., 2004), rats in the Surprise condition should show less generalisation between stimulus A and stimulus D, relative to the other three conditions.

3.7.2 Method

3.7.2.1 Subjects

Thirty-two experimentally naïve male Lister hooded rats (*Rattus norvegicus*) obtained from OLAC, Bicester, UK were maintained at 80% of their free feeding weights (mean: 395.6g; range: 363g-422g) by giving them a restricted quantity of food (Teklad laboratory diet, Harlan Teklad, Bicetser, Oxfordshire, UK) at the end of each day. All rats were housed in pairs in a colony room that was illuminated between 8:00 a.m. and 8:00 p.m. Each housing cage contained a single cardboard tube (18.0 cm length × 10.0 cm diameter) throughout the course of the experiment. Eight rats served as subjects in each of the four conditions (see Table 4 for an overview of the experimental design).

Table 4. *Experimental design of Experiment 9.*

Condition	Preexposure		Magazine Training	Conditioning	Test
	1st Four Days	2nd Four Days	Two Days	Three Days	Two Days
Baseline	A / - / - / D	D / - / - / A	+	A+	D
Surprise	A / B / - / D	D / - / B / A	+	A+	D
Sys_trans	A / B / C / D	D / C / B / A	+	A+	D
Scram_trans	A / C / B / D	D / B / C / A	+	A+	D

Note. A, B, C and D represent four separate tone stimuli of 1 kHz, 2 kHz, 3 kHz and 4 kHz, respectively. + denotes the delivery of a single food pellet.

3.7.2.2 Apparatus

Four standard operant chambers (23.0 cm length × 24.5 cm width × 21.0 cm height; Campden Instruments Ltd., Loughborough, England) housed in sound- and light-resistant cabinets were used. After rats had been placed in the operant chambers, the doors of the cabinets were closed for testing. The chambers were arranged in a 2 × 2 array, and each received local illumination from a single house light. Each chamber was equipped with a food well into which 45-mg of food pellets could be delivered. A transparent plastic flap, 6 cm high × 5 cm wide, hinged along the top of the food well opening, guarded access to the food well. A movement of this flap of, approximately, 2 mm was automatically recorded as a single response or food well entry. The floors of the chambers were constructed from stainless steel rods (with diameters of 5 mm and mounted 15 mm apart). A speaker mounted on the ceiling of each operant chamber was used to present the auditory stimuli A, B, C and D. The four, 10-s auditory stimuli were constant tones of 1 kHz, 2 kHz, 3 kHz and 4 kHz (produced by one audio generator; Campden Instruments Ltd., Model no. 258). These stimuli were presented at an intensity of, approximately, 75 dB (A weighting). A computer controlled the apparatus and recorded all food well entries.

3.7.2.3 Design and procedure

Preexposure phase

Stimulus preexposure lasted a total of eight days. Whether or not a rat received stimulus exposure on each of the eight days of stimulus preexposure was determined by whether the rat was in the Baseline, Surprise, Sys_trans, or Scram_trans condition (see Table 4). When a rat was scheduled to receive stimulus exposure, rats received 20, 10-s presentations of a single tone stimulus, separated by a 30-s inter-trial interval between the offset of one tone presentation and the onset of another. One preexposure session, therefore, lasted a total of 13 min 20-s. When a rat was not scheduled to receive stimulus exposure, they were simply placed into the operant chamber for a duration of 13 min 20-s.

Initially, rats were transferred from their home cages to the operant chambers. All rats in the Baseline condition received stimulus exposure on days 1, 4, 5 and 8; on days 2, 3, 6 and 7 these rats did not receive any stimulus exposure. Half of rats in the Baseline condition received stimulus exposure in the configuration A, -, -, D, D, -, -, A (1 – 8), while the other half of rats received stimulus exposure in the configuration D, -, -, A, A, -, -, D (1 – 8). ('-' indicates the absence of stimulus exposure). For half of rats in the Surprise condition, exposure was given to stimuli A, B, and D, and for the other half of rats, exposure was given to stimuli A, C, and D. For rats exposed to stimuli A, B, and D, half of these rats received stimulus exposure in the configuration A, B, -, D, D, -, B, A (1 – 8), while the other half of these rats received stimulus exposure in the configuration D, -, B, A, A, B, -, D (1 – 8). For rats exposed to stimuli A, C, and D, half of these rats received stimulus exposure in the configuration A, -, C, D, D, C, -, A (1 – 8), while the other half of these rats received stimulus exposure in the configuration D, C, -, A, A, -, C, D (1 – 8). Rats allocated to condition Sys_trans or condition Scram_trans received stimulus exposure on each of the 8 preexposure days. In condition Sys_trans, half of rats received stimulus exposure in the configuration A, B, C, D, D, C, B, A (1 – 8), while the other half of these rats received stimulus exposure in the configuration D, C, B, A, A, B, C, D (1 – 8). In condition Scram_trans, half of rats received stimulus exposure in the configuration A, C, B, D, D, B, C, A (1 – 8), while the other half of these rats received exposure in the configuration D, B, C, A, A, C, B, D (1 – 8).

Magazine training

Following the preexposure phase, over the next two days rats were trained to collect food pellets (Noyes Precision Pellets supplied by Sandown Chemicals Ltd, Hampton, England) from the food well. On the first day of training, the plastic flaps that guarded access to the food wells were fixed in a raised position to allow rats clear sight of, and easy access to, the food pellets. During the second day of training, the plastic flaps were lowered to their normal positions, and rats had to move the flaps to gain access to the food pellets. On each day of training, a total of 20 food pellets were delivered one at a time on a fixed-time 60-s schedule.

Conditioning and test

On the three days that followed magazine training, an appetitive response was conditioned to one of the two stimulus endpoints (i.e., A or D). Within each of the subconditions created through the previous counterbalancing operations employed during the preexposure phase, half of rats received appetitive conditioning to stimulus A, and half of rats received appetitive conditioning to stimulus D. On each day, rats received one session of appetitive conditioning in which they received 20, 10-s presentations of their scheduled tone stimulus, separated by a 30-s inter-trial interval. A single food pellet was delivered immediately after the offset of each 10-s tone presentation, meaning that each rat received a total of 20 food pellets per session.

Following appetitive conditioning, rats received two test days in which generalisation of the appetitive response to the opposite stimulus endpoint was assessed (A or D). Specifically, if an appetitive response had been conditioned to stimulus A, then these rats received a total of eight, 10-s nonreinforced presentations of stimulus D on each test day. In contrast, if an appetitive response had been conditioned to stimulus D, then these rats received a total of eight, 10-s nonreinforced presentations of stimulus A on each test day. Presentations of the eight, 10-s stimulus were separated by a 30-s inter-trial interval.

3.7.2.4 Measures

To assess appetitive conditioning, I compared the number of food well entries that rats made during presentations of the reinforced stimulus (CS) to the number of food well entries made during a 10-s pre-stimulus period (PCS). Successful appetitive conditioning was taken to reflect two observations: First, that the difference in the number of food well entries made during the PCS and CS was larger on day 3 of conditioning than on day 1 of conditioning. Second, that the number of food well entries made during the CS was significantly greater than the number of food well entries made during the PCS by day 3 of conditioning.

Generalisation of the conditioned appetitive response at test was taken to reflect the total number of food well entries made during the eight presentations of the test stimulus. For the purpose of analyses, these eight test trials were split into two equal blocks of four test trials.

3.7.3 Results

Figure 19 displays the overall results from appetitive conditioning (see Appendix 2, Table 8, for PCS and CS means split by condition). Inspection of this figure reveals that, overall, the number of food well entries made during both the PCS and CS declined across conditioning. While one might have expected the number of food well entries made during the CS to increase across conditioning, this decline is likely due to the large number of trials (i.e., 20) given on each day of conditioning. That is, across conditioning, rats became more targeted in their responding. Importantly, conditioning was restricted to three days so as not to undermine any effect of preexposure. What is critical, however, is that the overall difference between the number of food well entries made during the PCS and CS became larger as conditioning progressed. Moreover, rats made a far greater number of food well entries during the CS than during the PCS. ANOVA, with condition (Baseline, Surprise, Sys_trans or Scram_trans), day (1-3), and conditioning period (PCS or CS) as factors, revealed no effect of condition, $F(3, 28) = 1.72, p > .05$, a significant effect of day, $F(2, 56) = 3.87, p < .03$, and a significant effect of conditioning period, $F(1, 28) = 55.66, p < .001$. None of the interactions between these factors were significant (largest $F(3, 28) = 2.35, p > .05$). On day 3 of conditioning, the number of food well entries made during the CS was significantly greater than the number of food well entries made during the PCS (as assessed with a Bonferroni-corrected paired samples t-test, $t(31) = -5.68, p < .001$). I took this to be satisfactory evidence that by day 3 of conditioning, rats had acquired an appetitive response to the CS.

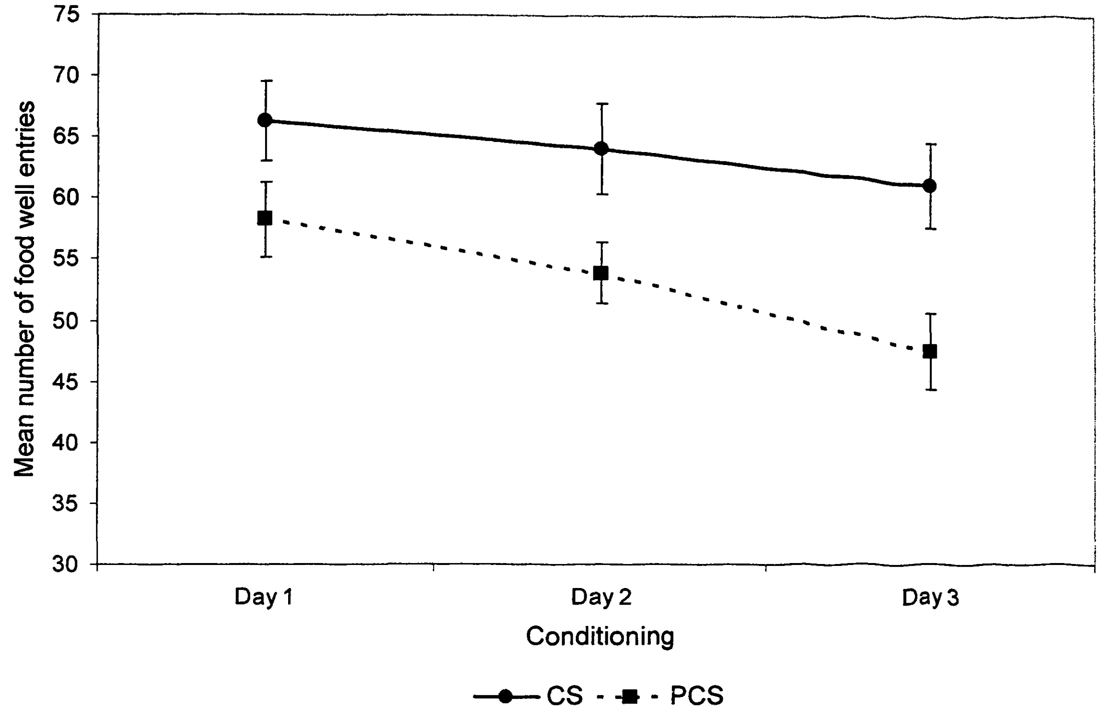

Figure 19. Appetitive conditioning in Experiment 9: overall mean number of food well entries across the three days of conditioning. Error bars indicate the standard error.

The results of principle interest are presented in Figure 20: for presentation purposes, the data are presented pooled over the two test days. Inspection of this figure reveals that while the number of food well entries made during the first four test trials (Block 1) differed little between preexposure conditions, marked differences were observed in the final four test trials (Block 2). Specifically, rats in condition Sys_trans and Scram_trans made fewer food well entries during block 2 of test than rats in either the Baseline condition or the Surprise condition. ANOVA, with condition (Baseline, Surprise, Sys_trans or Scram_trans), day (1-2), and block (1-2) as factors, revealed a main effect of day, $F(1, 28) = 36.13, p < .001$, and block, $F(1, 28) = 11.60$, $p < .003$. While the main effect of condition was found not to be significant, $F(3, 28) = 1.58$, $p = .22$, there was a significant interaction between block and condition, $F(3, 28) = 3.13$, $p < .05$. No other interactions were found to be significant (Fs<1). Critically, simple main effects revealed that while there was no effect of exposure condition in block 1 of test, $F(3, 112) = 1.14$, $p > .05$, there was a highly significant effect of exposure condition in block 2 of test, $F(3, 112) = 4.72$, $p < .004$. Collapsed across day, simple comparisons confirmed that in block 2 of test, rats in condition Sys_trans and Scram_trans made significantly fewer food well entries during presentations of the test stimulus than rats in both the Baseline condition and the Surprise condition (smallest $F(1, 112) = 6.21$, $p < .015$).

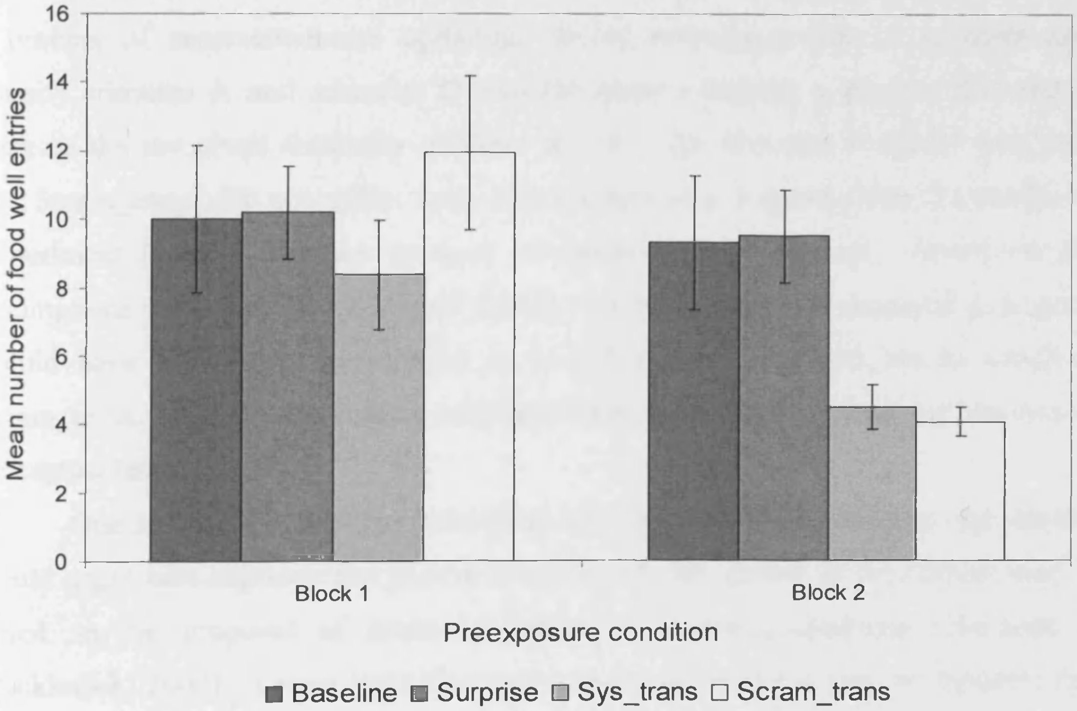

Figure 20. Results from the generalisation test of Experiment 9: mean number of food well entries collapsed across the two test days and split by block. Error bars indicate the standard error.

3.7.4 Discussion

The present findings indicate that rats that were preexposed to all four tone stimuli came to perceive stimulus A and stimulus D as more distinct than did rats that were preexposed to two or three of the four tone stimuli. Interestingly, the order in which the four stimuli were preexposed made little difference to rat's food well responding at test, and no significant differences were found between condition Sys_trans and condition Scram_trans. While acknowledging the fact that there are a number of important differences between this experiment and Experiment 6 (e.g., the fact that different stimuli were used), the current findings clearly sit in contrast to the findings of Experiment 6. Specifically, they provide no evidence that rats in the Surprise condition spontaneously classified stimulus A and stimulus D into different categories; an operation that should have decreased the effective similarity of these stimuli relative to the Baseline condition, at least. The present results, therefore, do not support the predictions of a surprise-driven category invention mechanism in rat spontaneous categorisation. Furthermore, the results also do not support the predictions made based on the assumption that transformational information, through a process of representational updating, should encourage rats to spontaneously classify stimulus A and stimulus D into the same category; a process that should increase the perceived similarity of these stimuli. The fact that condition Sys_trans and Scram_trans did not differ from one another also suggests that the results of Experiment 9 were not the product of categorical perception. Based on the assumptions of Newell and Bülthoff (2002), described earlier, categorical perception should have only been encouraged in condition Sys_trans and not in condition Scram_trans. Rather, the results of Experiment 9 appear to reflect an instance of perceptual learning.

One mechanism that has been proposed for perceptual learning, and which I would argue best captures the present results given the nature of the stimuli used, is based on the proposal of latent inhibition of common elements (McLaren & Mackintosh, 2000). Latent inhibition refers to the observation that preexposure to a stimulus will later retard subsequent conditioning to that stimulus (Lubow, 1989). To explain perceptual learning through latent inhibition, it has been proposed that when stimuli share common elements, latent inhibition will differentially affect the stimuli's common and unique features. As McLaren and Mackintosh note, "The argument rests on the seemingly plausible, even incontrovertible, assumption that the magnitude of any latent inhibition effect will be proportional to the amount of exposure to the stimulus or stimulus elements in question" (2000, p. 228). If one assumes, therefore, that the four tone stimuli presented in this experiment contain both unique elements (e.g., *a*, *b*, *c* and *d*) and common elements (e.g., *x*), then it is clear that rats in condition Sys_trans and Scram_trans will simply have received more preexposure to *x* than rats in the other two conditions. Consequently, when either stimulus A or stimulus D is later paired with food, it seems reasonable to assume that rats in condition Sys_trans and Scram_trans will be subject to greater latent inhibition of the common elements *x* than rats in either the Baseline or Surprise condition. Therefore, the appetitive conditioning will accrue preferentially to the unique elements of the conditioned stimulus (e.g., *a* or *d*) in condition Sys_trans and Scram_trans, at the expense of the shared elements (i.e., *x*). Thus, there will be less of a basis for generalisation of the appetitive response from, for example, stimulus A to stimulus D at test, which is driven by their shared elements (*x*). While it must be acknowledged that, based on this account, one would have expected to also see a lower level of generalisation in the Surprise condition, relative to the Baseline condition, it is possible that the lack of this difference might simply reflect the fact that this effect is quite small. However, this result suggests that the stated argument of McLaren and Mackintosh (2000, p. 228) needs to be refined. The present mechanism is favoured over, for example, the process of unitisation, given the nature of the stimuli used. Unitisation refers to the process of establishing a more veridical representation of the stimulus being sampled through the formation of associations between the elements that make up a stimulus. A tone stimulus has generally been regarded as a simple stimulus, meaning that most of its elements will be sampled on any given presentation. The impact of unitisation, therefore, will only be influential when dealing with complex stimuli, where different elements of a stimulus are sampled on each presentation, until a veridical representation has been formed (see McLaren & Mackintosh, 2000).

One reason why latent inhibition may have come to produce the present perceptual learning effect is due to the temporal dynamics of the preexposure regime employed (i.e., stimulus preexposure occurred across different days). A number of studies have found that spaced stimulus preexposure enhances latent inhibition, relative to massed stimulus preexposure (e.g., Schnur & Lubow, 1976); an observation which has been confirmed in modelling work (see McLaren & Mackintosh, 2000). A straightforward prediction from this work, therefore, is that the present perceptual learning effect should be attenuated following massed stimulus preexposure. What is more, as has been noted at various points in this chapter, increasing the temporal contiguity between stimulus presentations may encourage the formation of stronger excitatory associations between stimuli (see Hall, 1991; Honey

& Bateson, 1996). It is possible, then, that if rats' preexposure to the simple tone stimuli used here was massed together closer in time, this may also encourage a beneficial effect of transformational information. Irrespective of this theoretical analysis, one clear difference between Experiment 6 in humans and Experiment 9 in rats (apart from the nature of the stimuli used), regards the schedule of stimulus preexposure. While stimulus exposure in Experiment 6 was very much massed, stimulus exposure in Experiment 9 was not. Experiment 10, therefore, sought to test whether massed stimulus exposure would attenuate the perceptual learning effect found in Experiment 9. If it does, then these results would be somewhat more consistent with the results of Experiment 6, where participants' responding was found to be equivalent between the Baseline, Sys_trans, and Scram_trans conditions.

3.8 Experiment 10

3.8.1 Introduction

To this end, Experiment 10 employed only two of the four conditions of Experiment 9; namely, the Baseline condition and condition Sys_trans. These conditions were focused upon as they represent the minimal and maximal amounts of preexposure to the shared elements (*x*), and because it is assumed that any influence of transformational information, through a process of representational updating, should be most likely to be found following systematic presentations of the four tone stimuli. The basic design of the present experiment was the same as that of Experiment 9, with the exception that stimulus preexposure was now massed rather than spaced. That is, all scheduled tone stimuli were now presented on the same day, with preexposure occurring over a two day period.

3.8.2 Method

3.8.2.1 Subjects and apparatus

Sixteen experimentally naïve male Lister hooded rats (*Rattus norvegicus*) obtained from OLAC, Bicester, UK were maintained in exactly the same way as in Experiment 9 (mean free feeding weights: 482.5g; range: 442g-534g). Eight rats served as subjects in the Baseline condition and eight rats served as subjects in condition Sys_trans (see Table 5). Rats in the Baseline condition were preexposed only to stimuli A and D, and rats in condition Sys_trans were preexposed to all four tone stimuli in sequence (A, B, C then D). The apparatus used was that of Experiment 9.

Table 5. *Experimental design of Experiment 10 and 11.*

Condition	Preexposure		Magazine Training	Conditioning	Test
	Day 1	Day 2	Two Days	Three Days	Two Days
Baseline	A / - / - / D	A / - / - / D	+	A+	D
Sys_trans	A / B / C / D	A / B / C / D	+	A+	D

Note. A, B, C and D represent four separate tone stimuli of 1 kHz, 2 kHz, 3 kHz and 4 kHz, respectively. + denotes the delivery of a single food pellet.

3.8.2.2 Preexposure, magazine training, conditioning and test

Stimulus preexposure occurred over the first two days of the experiment. On each day of stimulus preexposure, rats were given nonreinforced presentations of the tones they were scheduled to receive over four sessions; each session was separated by, approximately, 1 hr. Within each session of stimulus exposure, rats received 20, 10-s tone presentations, separated by a 30-s inter-trial interval between the offset of one tone presentation and the onset of another. Half of rats in the Baseline condition received stimulus preexposure in the configuration A, -, -, D, A, -, -, D (1 – 4), and the other half of rats received stimulus preexposure in the configuration D, -, -, A, D, -, -, A (1 – 4). For rats in condition Sys_trans, half received stimulus preexposure in the configuration A, B, C, D, A, B, C, D (1 – 4), and half received stimulus preexposure in the configuration D, C, B, A, D, C, B, A (1 – 4). As for Experiment 9, when rats in the Baseline condition were not scheduled to receive presentations of a tone stimulus (i.e., on sessions 2 and 3 of each day), they were simply placed into the operant chamber for 13 min 20 s (i.e., the duration of stimulus exposure on a trial). Within each of the subconditions created by the previous counterbalancing operation, half of rats received appetitive conditioning to stimulus A and were presented with stimulus D at test, and half of rats received appetitive conditioning to stimulus D and were presented with stimulus A at test. Magazine training, conditioning and the generalisation test all proceeded in exactly the same manner as for Experiment 9.

3.8.3 Results

Figure 21 displays the overall results from appetitive conditioning (see Appendix 2, Table 8, for PCS and CS means split by condition). As for Experiment 9, inspection of this figure reveals, critically, that the difference in responding during the PCS and CS became larger as conditioning progressed. Moreover, rats made a far greater number of food well entries during the CS than during the PCS. ANOVA, with condition (Baseline or Sys_trans), day (1-3), and conditioning period (PCS or CS) as factors, revealed no effect of condition, $F(1, 14) = 2.37, p > .05$, a significant effect of day, $F(2, 28) = 5.81, p < .009$, and a significant effect of conditioning period, $F(1, 14) = 32.17, p < .001$. None of the interactions between these factors were significant (largest $F(2, 28) = 1.94, p > .05$). On day 3 of conditioning, the number of food well entries made during the CS was significantly greater than the number of food well entries made during the PCS (as assessed with a Bonferroni-corrected paired samples t-test, $t(15) = -4.29, p < .002$). Again, I took this to be satisfactory evidence that by day 3 of conditioning, rats had acquired an appetitive response to the CS.

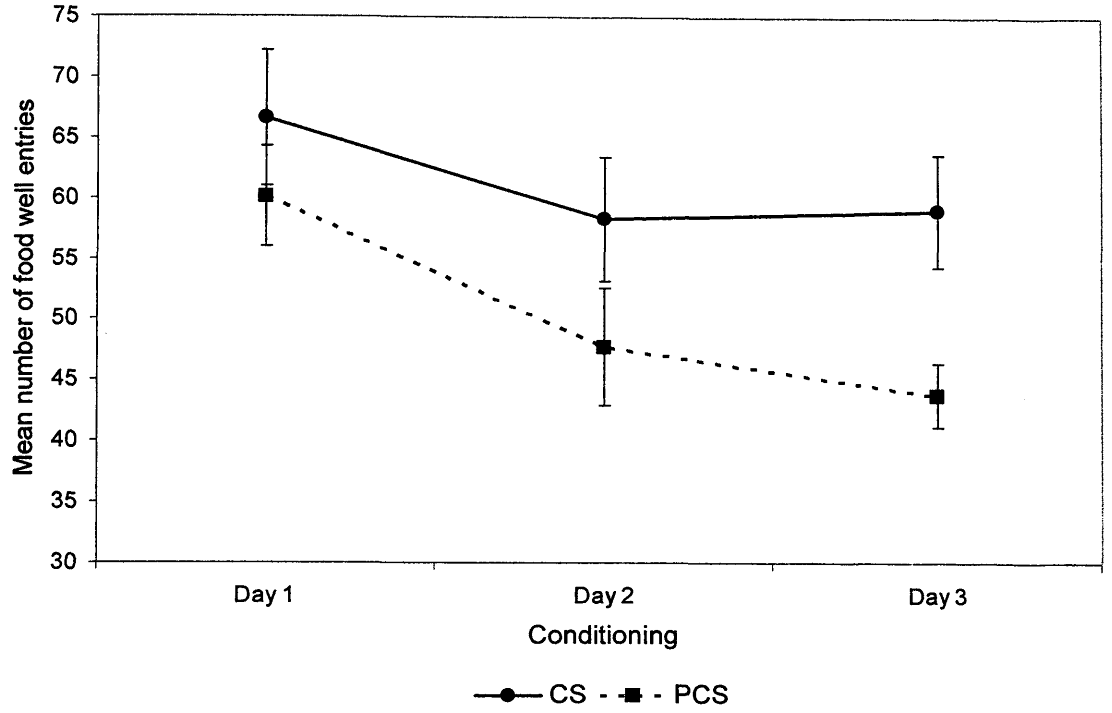

Figure 21. Appetitive conditioning in Experiment 10: overall mean number of food well entries across the three days of conditioning. Error bars indicate the standard error.

Concerning the results of principle interest (see Figure 22), one rat was removed from this analysis on the basis of being a clear statistical outlier, defined as being over two standard deviations away from the overall condition mean. This rat had served as a subject in the Baseline condition. Figure 22 displays the results of interest; again, for presentation purposes, the data are presented pooled over the two test days. Block 1 refers to the first four test trials, and block 2 refers to the final four test trials. Inspection of this figure reveals that while little difference existed between the two conditions in block 1 of test, over the final four trials in block 2 of test, a marked difference emerged. Interestingly, the pattern of results was in the opposite direction to those of Experiment 9. That is, rats in condition Sys_trans made a greater number of food well entries in block 2 of test than rats in the Baseline condition. ANOVA, with condition (Baseline or Sys_trans), day (1-2), and block (1-2) as factors, revealed significant main effects of day, $F(1, 13) = 6.95, p < .025$, and block, $F(1, 13) = 15.46, p < .003$, but no main effect of condition, $F(1, 13) = 1.25, p > .05$. Furthermore, none of the interactions between these factors were significant (largest $F(1, 13) = 1.65, p > .05$, which represents the block × condition interaction).

While the interaction between block and condition was found not to be significant in Experiment 10, when split by block, there is a clear trend in the data. To reiterate, while there is little difference in the number of food well entries made between the two conditions in block 1, in block 2, rats in condition Sys_trans produced a greater number of food well entries than rats in the Baseline condition. This trend in the data was explored using a Bonferroni corrected critical value of *p*<.025. Follow-up tests revealed that while the two conditions do not differ significantly from one another in block 1 (*F*<1), when collapsed over day, rats in condition Sys_trans produced a greater number of food well entries in block 2 than rats in the Baseline condition, and this difference closely approached significance at the corrected value (*F*(1, 13) = 5.51, *p* = .035).

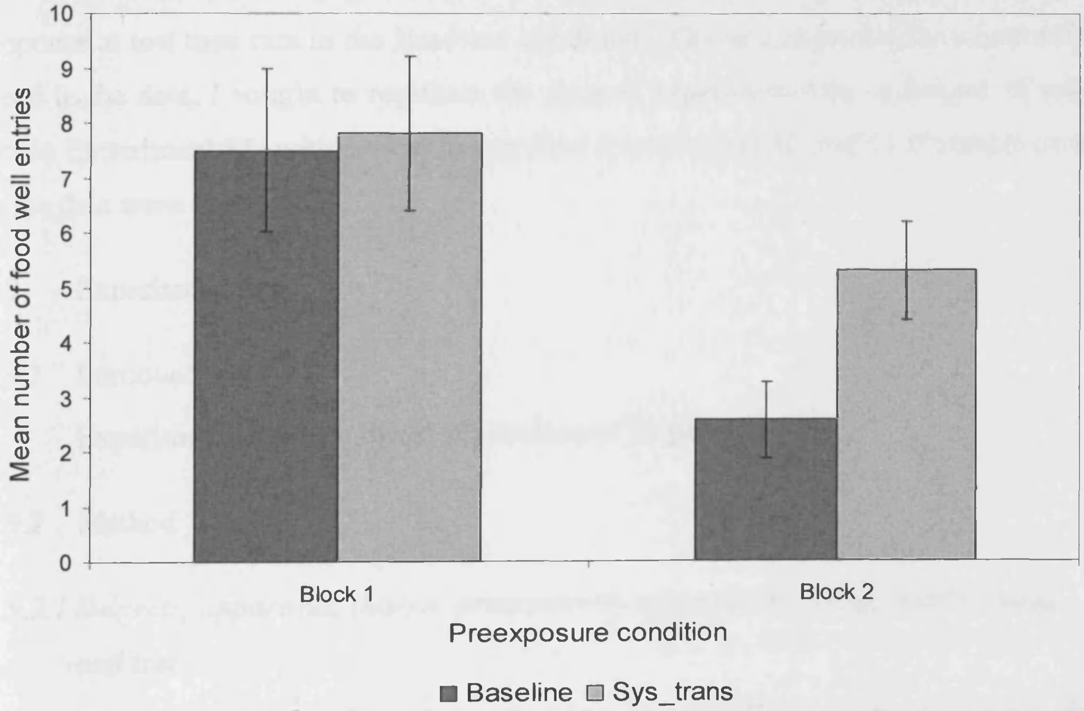

Figure 22. Results from the generalisation test of Experiment 10: mean number of food well entries collapsed across the two test days and split by block. Error bars indicate the standard error.

3.8.4 Discussion

The results of Experiment 10 show a pattern of results opposite to that of Experiment 9: rats in condition Sys_trans made more magazine entries in block 2 of test than did rats in the Baseline condition, although the two conditions did not differ significantly from each other at any point. In this regard, therefore, the results are more similar to the pattern of results found in Experiment 6, suggesting some level of consistency in incidental categorisation between humans and rats. These results, therefore, provide some support for the idea that the schedule of stimulus preexposure influences perceptual learning (McLaren & Mackintosh, 2000; Schnur & Lubow, 1976). Specifically, massing stimulus preexposure appears to reduce the influence of perceptual learning.

Of particular interest in the results of Experiment 10 is the trend that exists in the data: while the two conditions showed little difference in number of food well entries during block 1, during block 2, rats in condition Sys_trans produced a greater number of food well entries than rats in the Baseline condition. This finding suggests that rats in condition Sys_trans showed greater generalisation of the appetitive response at test than rats in the Baseline condition. Given this particularly interesting trend in the data, I sought to replicate the present experiment with a further 16 naïve rats in Experiment 11, with a view to combine Experiments 10 and 11 if similar trends in the data were observed.

3.9 Experiment 11

3.9.1 Introduction

Experiment 11 was a direct replication of Experiment 10.

3.9.2 Method

3.9.2.1 Subjects, apparatus, design, preexposure, magazine training, conditioning, and test

Sixteen experimentally naïve male Lister hooded (*Rattus norvegicus*) rats (free feeding weights mean: 390.4g; range: 359g-412g) that came from the same supplier and were maintained in the same way as those used in Experiments 9 and 10. Again, eight subjects served in each condition. All aspects of the apparatus, design and procedure were identical to Experiment 10 (see Table 5).

3.9.3 Results and discussion

Figure 23 displays the overall results from appetitive conditioning (see Appendix 2, Table 8, for PCS and CS means split by condition). Once again, the difference in responding during the PCS and CS was larger on day 3 of conditioning than on day 1 of conditioning. Moreover, rats made a far greater number of food well entries during the CS than during the PCS. ANOVA, with condition (Baseline or Sys_trans), day (1-3), and conditioning period (PCS or CS) as factors, revealed no effect of condition, $F<1$, no effect of day, $F(2, 28) = 2.24$, $p > .05$, and a significant effect of conditioning period, $F(1, 14) = 74.34$, $p < .001$. A significant interaction between day and conditioning period, $F(2, 28) = 3.82$, $p < .05$, was also found (no other interactions were significant). Simple main effects revealed that the number of food well entries made during the CS was significantly greater than the number of food well entries made during the PCS at each day of conditioning (smallest $F(1, 14) = 12.97$, $p < .003$). It is clear, however, that this difference was most pronounced on day 3 of conditioning. As for Experiments 9 and 10, I took this to be satisfactory evidence that by day 3 of conditioning, rats had acquired an appetitive response to the CS.

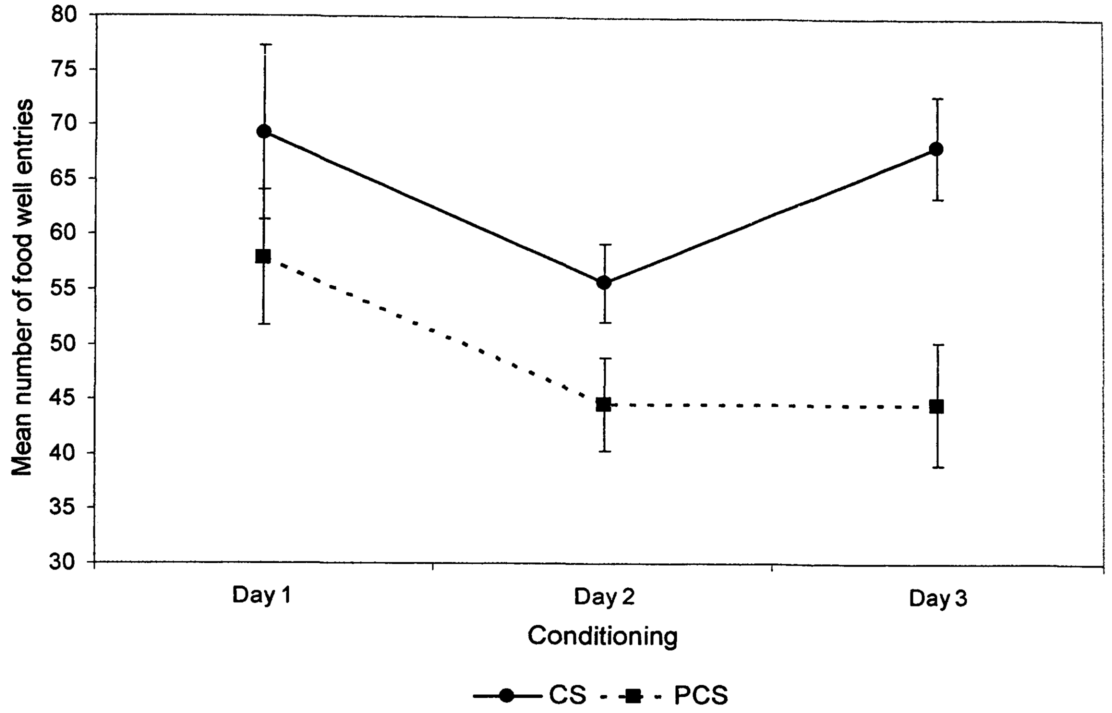

Figure 23. Appetitive conditioning in Experiment 11: mean number of food well entries across the three days of conditioning. Error bars indicate the standard error.

Focusing now on the results of principle interest displayed in Figure 24 (presented in the same way as for Experiment 10), it is clear that the results confirm the findings of Experiment 10. Indeed, the pattern of results is near identical. Once again, therefore, rats in condition Sys_trans showed a higher level of responding in block 2 of test than did rats in the Baseline condition. ANOVA, with day (1-2), block (1-2), and condition (Baseline or Sys_trans) as factors, revealed significant main effects of day, $F(1, 14) = 13.66, p < .003$, and block, $F(1, 14) = 6.14, p < .03$, but no effect of condition, $F<1$. Moreover, none of the interactions between these factors were significant, however the day × block interaction closely approached significance (largest $F(1, 14) = 4.16, p = .06$).

Given the same interesting trend as for Experiment 10, a Bonferroni corrected critical value of $p<.025$ was again employed to further explore the data. Follow-up tests revealed that rats in the Baseline condition and condition Sys_trans did not differ significantly in their level of responding either on block 1 ($F<1$) or block 2 of test ($F(1, 14) = 2.07, p > .025$). To increase the power of any conclusions, the results of Experiment 10 and 11 were combined, and statistical analyses run on this expanded data set.

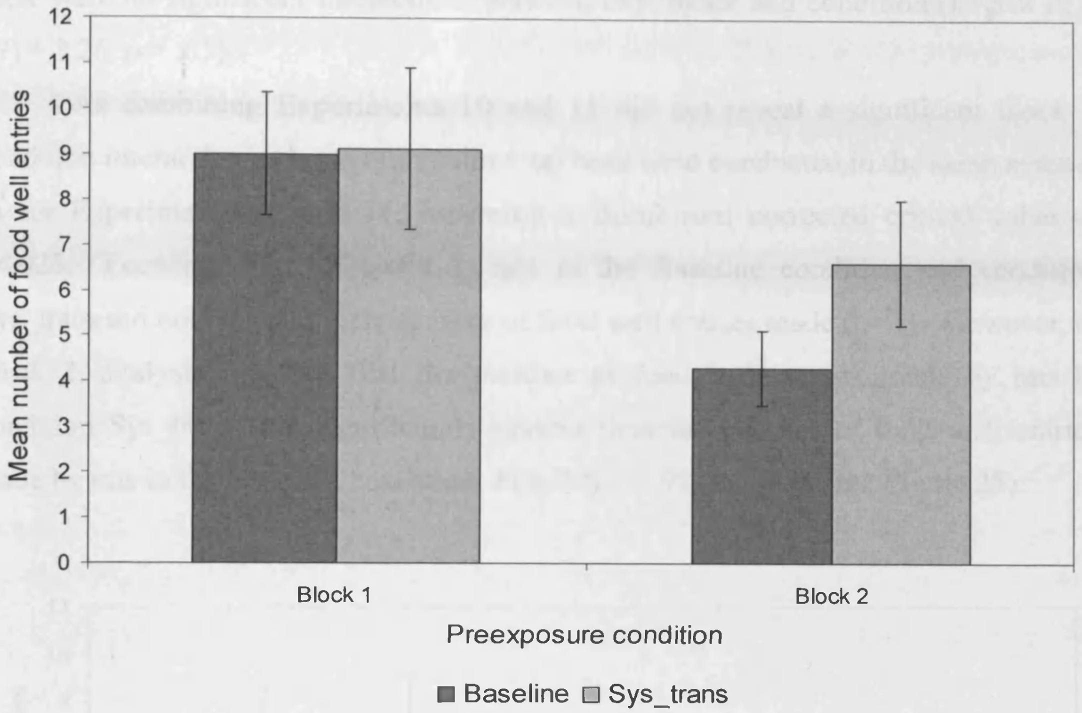

Figure 24. Results from the generalisation test of Experiment 11: mean number of food well entries collapsed across the two test days and split by block. Error bars indicate the standard error.

3.10 Combining Experiments 10 and 11

The method chosen to combine Experiments 10 and 11 involved introducing 'Experiment' as a second between-subjects variable in the ANOVA[11]. If no significant differences are found with respect to this factor, then further conclusions will be drawn. To this end, ANOVA, with experiment (Experiment 10 or Experiment 11), condition (Baseline or Sys_trans), day (1-2), and block (1-2) as factors, revealed that there was no main effect of experiment, *F*(1, 27) = 2.43, *p* > .05, and that there were no significant interactions between experiment and any other factor (largest *F*(1, 27) = 1.33, *p* > .05). Collapsing across experiment, therefore, ANOVA confirmed significant main effects of day, *F*(1, 27) = 19.65, *p* < .001, and block, *F*(1, 27) = 17.29, *p* < .001, but no main effect of condition, *F*(1, 27) = 2.13, *p* > .05. Moreover,

[11] The meta-analytic procedures described by Rosenthal (1991) produced results that were entirely consistent with the findings presented below, where 'Experiment' was introduced as a second between-subjects variable.

there were no significant interactions between day, block and condition (largest $F(1, 27) = 2.26, p > .05$).

As combining Experiments 10 and 11 did not reveal a significant block × condition interaction as hoped for, follow-up tests were conducted in the same manner as for Experiments 10 and 11, assuming a Bonferroni corrected critical value of $p<.025$. Focusing first on block 1, rats in the Baseline condition and condition Sys_trans did not differ in their number of food well entries made ($F<1$). However, in block 2, analysis revealed that the number of food well entries made by rats in condition Sys_trans was significantly greater than the number of food well entries made by rats in the Baseline condition, $F(1, 29) = 5.97, p < .025$ (see Figure 25).

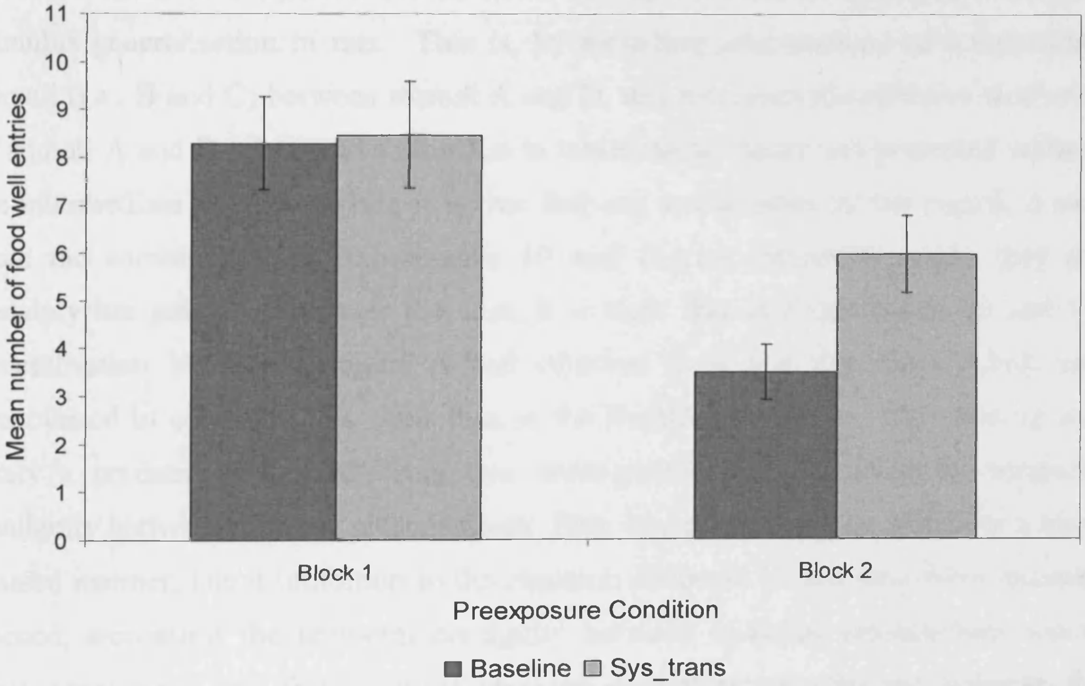

Figure 25. Results from combining the generalisation tests of Experiment 10 and Experiment 11: mean number of food well entries collapsed across the two test days and split by block. Error bars indicate the standard error.

3.10.1 Conclusions

Overall, the pattern of results found in Experiments 10 and 11 sits in contrast to the pattern of results found in Experiment 9. While the pattern of results found in Experiment 9 appeared to document an instance of perceptual learning, the results of

Experiment 10 and 11 documented an attenuation of this perceptual effect. These findings are consistent with those of Schnur and Lubow (1976), and the predictions of McLaren and Mackintosh (2000), that more massed stimulus exposure attenuates the influence of perceptual learning. I would argue, therefore, that the results of Experiments 10 and 11 indirectly favour an account of the perceptual learning effect seen in Experiment 9 based on latent inhibition to the common elements (see McLaren & Mackintosh, 2000).

Of particular interest is the fact that while combining Experiments 10 and 11 did not reveal the significant block × condition interaction hoped for, follow-up tests confirmed that rats in condition Sys_trans made significantly more food well entries during block 2 than rats in the Baseline condition. This important finding documents the first evidence of a facilitative influence of transformational information on later stimulus generalisation in rats. That is, by including presentations of intermediate stimuli (i.e., B and C) between stimuli A and D, this increased the effective similarity of stimuli A and D relative to a situation in which these stimuli are presented without the intermediate stimuli. While it is true that any conclusions in this regard, drawn from the combination of Experiments 10 and 11, are inherently weak, they are certainly intriguing. Whatever the case, it is clear that in Experiments 10 and 11, generalisation between stimulus A and stimulus D at test was more robust and pronounced in condition Sys_trans than in the Baseline condition. This finding was likely a product of the following two consequences of increasing the temporal contiguity between stimulus presentations: first, by preexposing the stimuli in a more massed manner, latent inhibition to the common elements would have been reduced. Second, increasing the temporal contiguity between stimulus presentations would have encouraged the formation of stronger excitatory associations between the stimuli. This would have been particularly prominent in condition Sys_trans, where preexposure to each stimulus was only separated by, approximately, an hour. One further possibility is that rats in condition Sys_trans engaged in a process of representational updating, whereby the discrete stimulus presentations were integrated into some single representation (e.g., "Tone") and 'classified together' (cf. Ryan & Lea, 1990). This latter account of the results is not favoured, however, as the process of representational updating should not be particularly affected by the schedule of stimulus exposure. Indeed, in the natural world, representational updating will likely be a fairly slow process, occurring over a long period of time. That is, if chicks do integrate discrete "snapshots" of their mother hen into a single representation (Bateson, 1973), these different "snapshots" will be accumulated over many days and months. Consequently, if rats do engage in representational updating, then there is no reason to believe that they should not have engaged in such a process in Experiment 9 (which would have overridden any extra influence of latent inhibition to the common elements). Rather, the results of Experiments 10 and 11 appear to reflect an instance of sensory-preconditioning (Hall, 1991).

3.11 General Discussion

The findings of Chapter 3 highlight a number of interesting consequences that different conditions of stimulus preexposure can have on later stimulus generalisation. The interesting feature of the research presented here is that, rather than simply assessing the influence that stimulus preexposure has *per se* (i.e., comparing a situation in which some form of stimulus preexposure is given to a situation in which no preexposure is given), the experiments have been focused towards understanding how perceived structure, contained within the distributional properties of a set of stimuli, influences stimulus similarity. This approach was taken so as to be able to focus on categorisation that was truly incidental and spontaneous, rather than on categorisation that was guided either by some explicit instruction to categorise or through reinforcement (feedback). In particular, I was interested in trying to better understand what aspects of perceived structure within a set of distributed stimuli will come to influence whether those stimuli are incidentally 'classified together' into the same category, or 'classified apart' into different categories. To assess this, the influence of four different conditions of stimulus exposure on later stimulus generalisation was assessed.

Following on from the work presented in Chapter 2, in the first three experiments I focused on human incidental categorisation. In Experiment 6, I found evidence to support the view that incidental categorisation in humans is guided by a surprise-driven category invention mechanism (see Clapper & Bower, 1994, 2002; Love et al., 2004). While it is possible that the temporal dynamics of stimulus preexposure was influential in generating these results, this was ruled out in Experiment 7. Rather, based on the similarity structure of the complex, naturalistic morph stimuli presented, I would argue that participants in the Surprise condition came to spontaneously classify the object category endpoints (i.e., stimuli A and F) into different categories. As a result of this incidental categorisation, participants in the Surprise condition reported a lower level of property generalisation between stimuli A and F, relative to participants in the other three conditions.

Interestingly, no evidence was found to support the prediction that transformational knowledge should enhance generalisation between two distinct, but similar stimuli, relative to a situation in which no transformational knowledge existed (see Zaki & Homa, 1999). Indeed, the results of Experiment 8 showed that, under certain circumstances, transformational knowledge can actually lead to a reduction in the perceived similarity of stimuli A and F, by increasing the temporal spacing between presentations of these stimuli. That is, Experiment 8 found that when stimuli A and F were preexposed in a manner that was highly temporally contiguous (i.e., stimulus F followed immediately after stimulus A, for example), participants perceived these stimuli as significantly more similar than participants in a condition where presentation of stimuli A and F was separated by a temporal delay. This was true whether stimuli A and F were separated by a simple fixation cross, or by transformational knowledge (i.e., the intermediate, transformational steps that resulted from transforming stimulus A into stimulus F). This finding is reminiscent of work by Pothos, Hahn and Prat-Sala (2008): specifically, they found that for items about which participants had prior knowledge, a slow transformation from one object (e.g., A) to a different object (e.g., B) can result in participants viewing A and B as less similar than in a situation in which an abrupt change occurs between these two stimuli. They explain their results in terms of psychological essentialism (see Malt, 1990; Medin & Ortony, 1989; Putnam, 1975; Rips, Blok, & Newman, 2006), such that participants who were told of a slow transformational change inferred something about the evolutionary origin of stimulus B, in which the essence of stimulus B has been altered through evolutionary pressures. This change in essence was not inferred by those participants told of an abrupt transformational change, and consequently, the effective similarity of stimulus A and stimulus B remained high. However, this view offers no reason as to why incorporating a temporal delay between presentations of stimuli A and F should reduce stimulus similarity compared to when stimuli A and F are presented in a temporally contiguous fashion.

More broadly, the results of Experiments 6 and 7 build on the work of Chapter 2 (see also Gureckis & Goldstone, 2008) in showing that within-category similarity structure is very important in human spontaneous categorisation. As has been noted, Murphy and Medin (1985) have argued that "we categorise not on the basis of a similarity cluster, but on the basis of selecting the concept that best explains the instance to be categorized" (Hampton, 2001, p. 16). Given the naturalistic stimuli that were employed in Experiments 6 – 8, it is important to appreciate that participants would have had some prior knowledge regarding the category of the presented stimuli as a whole; although not specifically about the individual stimuli *per se*. However, there is no reason to believe that the amount of prior knowledge that participants had differed between the different preexposure conditions. Consequently, given the four preexposure conditions, it seems that the only basis by which participants in the Surprise condition could have come to show a reduced amount of property generalisation between stimuli A and F, relative to participants in the other three conditions, was on the basis of one similarity cluster (containing, for example, stimuli A, B and C) being spontaneously classified as distinct from a second cluster (containing, for example, stimulus F). This result supports the intuitive notion that the clustering together of similar stimuli provides an important mechanism for human spontaneous categorisation (Hampton, 2001). To provide direct support for this claim, future research could also look to assess the amount of property generalisation (or perceived similarity) between the stimuli that form the cluster of highly similar stimuli. Based on the view outlined above, one would expect to see a greater amount of property generalisation between stimuli A and C in the Surprise condition than in the other three conditions (if the cluster of highly similar stimuli was formed from stimuli A, B and C, and the distinct cluster contained stimulus F).

One particularly interesting finding from Experiments 6 – 8 is that human participants do not appear inclined to incidentally 'classify apart' two distinct, but similar stimuli (e.g., stimuli A and F). That is, participants in the Baseline condition showed no evidence of classifying stimuli A and F into different categories. Rather, the incidental classification of two distinct, but similar stimuli is driven by the existence of other stimuli that are highly similar to one of the two stimuli, allowing for certain norms to be developed around the highly similar stimuli. Consequently, it is only through the creation of these norms that perceived discontinuity within the stimulus set becomes meaningful. The findings of Experiment 6 and 7, therefore, fit nicely with an account of human spontaneous categorisation based on a surprise-driven category invention mechanism, which operates on stimulus similarity structure (see Clapper & Bower, 1994, 2002; Love et al., 2004, see also, Anderson, 1991).

In conclusion, while it is true that prior knowledge increases the likelihood with which people come to identify a specific category structure (e.g., Clapper, 2007; Spalding & Murphy, 1996), the present results, and those of Chapter 2, clearly document that category structure can readily be imposed on a set of stimuli based purely on stimulus similarity structure. The results of Experiments 6 – 8, therefore, strongly suggest that perceived discontinuities within our environment, based on stimulus similarity, are an important influence in guiding human spontaneous categorisation (Rosch & Mervis, 1975).

While it is difficult to draw direct comparisons between Experiments 6 – 8 in human participants and Experiments 9 – 11 in rats, what is clear is that, in a formally equivalent design, rats showed a qualitatively different pattern of generalisation behaviour at test than did the human participants. These results, therefore, suggest a more primary role for the classifier in categorisation. Specifically, in Experiment 9, rats in condition Sys_trans and Scram_trans showed a reduced amount of generalisation of the appetitive conditioned response from, for example, stimulus A to stimulus D, relative to rats in the Baseline and Surprise conditions. I argued that this apparent perceptual learning effect was most likely the result of stronger latent inhibition to the common cues in conditions Sys_trans and Scram_trans, brought about by the extra amount of stimulus preexposure in these conditions. This account was strengthened by the results of Experiment 10 and 11, which showed that more massed stimulus presentation brought about a reversal in the aforementioned pattern of results. Specifically, in Experiments 10 and 11, rats in condition Sys_trans showed an increased amount of generalisation of the appetitive conditioned response from, for example, stimulus A to stimulus D, relative to rats in the Baseline condition. Despite these latter findings, I would argue that, overall, the results of Experiments 9 – 11 show little evidence to support the view that rats engaged in incidental, spontaneous categorisation.

Of course, a number of important differences between Experiments 6 – 8 and Experiments 9 – 11 may have contributed to the contrasting pattern of results found for humans and rats. First, whereas the stimuli presented to the human participants were complex, naturalistic objects, the stimuli presented to the rats were simple tone stimuli. Perhaps the tone stimuli were not distributed in a manner that was most likely to promote perceived discontinuity between the highly similar set of stimuli and the dissimilar stimulus in the Surprise condition of Experiment 9. Furthermore, it is worth noting that in the Surprise, Sys_trans and Scram_trans conditions, the number of stimuli preexposed to the human participants was greater than the number of stimuli preexposed to the rats; again, this may have reduced the perceived structure contained within the presented tone stimuli. What is more, stimulus preexposure was rather different in the human experiments than in the rat experiments (in terms of exposure schedule, at least).

While the above factors may have played a contributory role in producing the qualitatively different pattern of results found for humans and rats, one further possibility is that rats may simply not possess the required level of cognitive flexibility to spontaneously appreciate the similarity-based relationships that exist between stimuli. Indeed, a number of authors, for example Chater and Heyes (1994), have proposed that there exists no evidence to support the view that nonhuman animals engage in categorisation in a manner that is qualitatively similar to humans. If correct, then it is hardly surprising that rats showed a qualitatively different pattern of generalisation behaviour to humans in the experiments reported in this chapter. However, more recent connectionist analysis (e.g., Honey & Ward-Robinson, 2002) has shown that simple associative processes (albeit ones applied in a three-layer network) should be capable of affording true stimulus grouping behaviour in nonhuman animals, and experimental results have supported this (e.g., Honey & Watt, 1998, 1999). Moreover, this connectionist analysis, based on basic associative processes, supports a view of nonhuman categorisation that is far more flexible than some have assumed (e.g., Chater & Heyes, 1994). Consequently, in the final experimental chapter of this thesis (Chapter 4), I investigate whether rats exhibit another important aspect of human categorization; namely, stimulus cross-classification. As such, the role of the classifier in affording this complex form of categorisation behaviour was assessed.

Chapter 4

Cross-classification in rats

4 Introduction

As postulated at the end of Chapter 3, one possible reason why rats showed a qualitatively different pattern of generalisation behaviour to humans is due to a lack of required cognitive flexibility on the part of rats, which facilitates incidental categorisation. Premack, for example, has suggested that "only primates may sort the world, i.e., divide it into its indeterminately many classes" (1976, p. 215). Chater and Heyes have gone one step further, arguing for what can be seen as a qualitative distinction between human and nonhuman animal categorisation:

> "the significance of the distinction between symbolic labelling and association is that the same set of exemplars can be labelled by many different labels (so that, for example, a given pair of exemplars can be represented as both being instances of ANIMAL, DOG and FURRY, but as differing regarding FIERCE) whereas association between exemplars is merely present or absent. Therefore, while it is possible for different labels to capture many different classifications, which may cross-classify or be arranged in hierarchies, associations can only produce a single partition of exemplars into two or more disjoint sets" (1994, p. 216).

To recap, the results from the human studies of Chapter 3 showed that human participants in the Surprise condition – which received preexposure to three highly similar stimuli and one distinct stimulus (e.g., A, B, C and F) – showed a lower level of later property generalisation between stimuli A and F than participants that either received preexposure only to stimuli A and F, or to stimuli A, B, C, D, E and F. I argued that this result could be explained by assuming that the similarity structure of the Surprise condition encouraged participants to spontaneously classify the three highly similar stimuli (e.g., A, B and C) into a different category from the distinct stimulus (e.g., F), on the basis of surprise-driven category invention mechanism (see Clapper & Bower, 1994, 2002; Love et al., 2004). This pattern of results was not observed in rats; albeit using a very different set of stimuli.

While driven by similarity, the above pattern of assumed incidental classification in humans requires a certain level of cognitive flexibility. Specifically, in the human and rat experiments of Chapter 3, given the high level of within-category similarity, two levels of stimulus classification were possible: First, at the presumed basic level, which would encompass all of the similar stimuli into a single category, and second, at the subordinate level, which would result in further divisions of the basic level of categorisation. If one assumes that the nonreinforced stimulus exposure given in Chapter 3 encouraged the formation of associations between the stored representations of the similar stimuli (see McLaren & Mackintosh, 2000), such that they became connected with one another, then based on the arguments of Chater and Heyes (1994), only humans would have the cognitive flexibility to impose a further, subordinate level form of classification (see Gureckis & Goldstone, 2008). That is, it is supposed that rats may be incapable of treating the physically similar stimuli A, B, and D equivalently in one set of conditions, and differently in a second set of conditions.

The results of Chapter 3, therefore, beg the question of the flexibility of stimulus classification in rats. As documented in the previous chapter of this thesis, the classification of stimuli into the same category has been associated with an increase in stimulus similarity, whereas the classification of stimuli into different categories has been associated with a decrease in stimulus similarity (Harnad, 1987). In rats, this change in stimulus similarity through stimulus classification has been highlighted in demonstrations of the acquired equivalence and distinctiveness of cues. To recapitulate, Honey and Watt (1998, 1999), for example, gave rats appetitive training in which four compounds were paired with food (AX, BX, CY & DY), and four were paired with no food (CX, DX, AY & BY). Following training, A was paired with footshock and C was not. They found that this revaluation treatment resulted in B eliciting greater generalised fear than D. Given the fact that A shares no more common elements with B than it does with D, these results suggest that the initial appetitive training modified the effective similarity of the stimuli, such that A and B were seen as similar, whereas A and D were not. One interpretation of the results reported by Honey and Watt (1998, 1999) assumes that when similar compounds (e.g., AX & BX) are followed by the same outcome (e.g., food), their components (i.e., A, B & X) come to address a shared configural unit within a connectionist network (see Allman et al., 2004; Honey & Ward-Robinson, 2002). According to this form of analysis, the appetitive training stage from Honey and Watt (1998, 1999) should result in four such configural units of the following form: *ABX*, *CDX*, *ABY*, and *CDY*. Under these conditions, therefore, when A was later paired with shock, *ABX* and /or *ABY* should become active and linked to a representation of shock. Consequently, presentations of B will be more likely to provoke fear than D, as B will activate the configural units *ABX* and *ABY*.

The above presented work, and other research on the acquired equivalence and distinctiveness of cues, is particularly interesting as it challenges the most widely accepted account of stimulus generalisation, which is based upon the suggestion that stimuli activate sets of elements, and that while some of these elements might be uniquely activated by a particular stimulus presented during training, other elements will be commonly activated by both the training and test stimuli. According to this account, therefore, similarity is *fixed* between any two stimuli at a given point in time: similarity and generalisation both simply reflecting the proportion of common elements that the two stimuli activate (e.g., Atkinson & Estes, 1963; Pearce, 1994; see also, McLaren & Mackintosh, 2000, 2002). The connectionist approach outlined above, which is based on a process of configural grouping, makes an intriguing, yet straightforward prediction about the flexibility of rats' classification behaviour, which has not been the subject of investigation. This prediction concerns the possibility that rats might be capable of forming groupings that allow for the cross-classification of a given set of stimuli (e.g., A, B, C and D): for example, grouping A with B, and, C with D in some conditions, while grouping A with D, and, B with C in others. Based upon the rationale and experimental design outlined in the next paragraph, I examined this prediction in two experiments.

Imagine that a rat is given the following set of appetitive training trials: AX and BX are paired with food, CX and DX are paired with no food; AY and DY are paired with food, and BY and CY are paired with no food (see Table 6). According to the theoretical analysis described above, this training should result in the formation of the following four configural units: *ABX*, *CDX*, *ADY* and *BCY*. That is, A is grouped with B, and C is grouped with D, when these stimuli are presented with X, whereas, A is grouped with D, and B is grouped with C, when these same stimuli are presented with Y. Under these circumstances, subsequent aversive trials in which A, for example, is paired with shock, and C is paired with no shock, should result in *ABX* and *ADY* (but not *CDX* or *BCY*) becoming linked to shock. After such revaluation, it follows that B should be no more likely to elicit fear than D, as both B and D can activate a configural unit linked to shock (namely, *ABX* & *ADY*, respectively).

However, presentation of BX will elicit more fear than DX if dual activation of a single hidden unit that is linked to shock provokes more fear than does singly activating two hidden units linked to shock (see Allman et al., 2004). Specifically, whereas BX provides two sources of activation to hidden unit *ABX* (that is linked to shock), and a single source of activation to two units that are linked to no shock (*CDX & BCY*), DX provides two sources of activation to hidden unit *CDX* (that is not linked to shock), and a single source of activation to two units that are linked to shock (*ABX & ADY*). For the same reasons, DY should elicit greater fear than BY: briefly, DY provides dual input to hidden unit *ADY* (that is linked to shock), and BY provides dual input to *BCY* (that is not linked to shock).

Table 6. *Experimental designs for Experiments 12 and 13.*

Experiment 12			
Appetitive Training		Revaluation	Tests
AX→ food	AY→ food		
BX→ food	BY→ no food	A→ shock	1. BX & DX
CX→ no food	CY→ no food	C→ no shock	2. BY & DY
DX→ no food	DY→ food		
Experiment 13			
Appetitive Training		Revaluation	Tests
AX→ food	AY→ food		
BX→ food	BY→ no food	B→ shock	1. AX & CX
CX→ no food	CY→ no food	D→ no shock	2. AY & CY
DX→ no food	DY→ food		

Note. A, B, C, and D refer to four different wallpapered environments in which rats were placed; X and Y refer to two different auditory stimuli. Food denotes the delivery of a single food pellet, whereas no food denotes its absence. Shock refers to the delivery of footshock, and no shock refers to its absence.

If the pattern of results predicted above were observed (i.e., stimulus generalisation was modulated by 'context' X and Y), then it would represent an interesting observation in its own right. Moreover, these results would also provide further support for a connectionist analysis of the acquired equivalence and distinctiveness of cues. Finally, such contextual modulation of stimulus generalisation in rats would resonate with the flexible stimulus classification documented in humans, and the observation that human similarity judgements are influenced by the context in which those judgements are made (Barsalou, 1982; Medin et al., 1993; Tversky & Gati, 1978).

Experiment 12 used the design that was outlined earlier to assess the prediction that contextual modulation of similarity can be observed in rats. Experiment 13 used a variant of this design to both extend the generality of the results of interest, and to contrast two theoretical interpretations for them.

4.1 Experiment 12

4.1.1 Introduction

The experimental design used is summarised in the upper rows of Table 6. During the first stage of appetitive conditioning, rats were placed in four, visually distinct experimental chambers (A, B, C & D) in which two auditory stimuli could be presented (X or Y). Four of the resulting compounds (AX, BX, AY and DY) were paired with one outcome (food in the example presented in Table 6), and the remaining four compounds (CX, DX, BY and CY) were paired with a second outcome (e.g., no food). All rats then received aversive conditioning in which A was paired with footshock and C was not. Finally, rats received test trials in which X and Y were presented in conjunction with placement in B and D. If rats had come to group the stimuli in the manner anticipated by the proposed connectionist analysis – that is, through their capacity to activate different configural units (*ABX*, *CDX*, *ADY*, *BCY*) – then *ABX* and /or *ADY* should have become linked to shock during pairings of A with shock. Consequently, rats should show greater generalised fear to the compound BX than to the compound DX, and similarly, greater fear to DY than to BY.

4.1.2 Method

4.1.2.1 Subjects

Sixteen experimentally naïve male Lister hooded rats (*Rattus norvegicus*) obtained from OLAC, Bicester, UK were maintained at 80% of their free feeding weights (mean: 397g; range: 360-423g) by giving them a restricted quantity of food (Teklad laboratory diet, Harlan Teklad, Bicester, Oxfordshire, UK) at the end of each day. All rats were housed in pairs in a colony room that was illuminated between

8:00 a.m. and 8:00 p.m. Each housing cage contained a single cardboard tube (18.0 cm length × 10.0 cm diameter) throughout the course of the experiment. All procedures commenced at 2:00 p.m.

4.1.2.2 Apparatus

Four standard operant chambers (23.0 cm length × 24.5 cm width × 21.0 cm height; Campden Instruments Ltd., Loughborough, England) housed in sound- and light-resistant cabinets were used. The doors of the cabinets were left open throughout the experiment to allow the rats' behaviour to be video recorded (during the final test) using a Panasonic movie camera (model number: NV-M40). The chambers were arranged in a 2 × 2 array, and each received local illumination from a single house light and ambient illumination from an overhead strip-light on the ceiling of the experimental room. The walls and ceiling of each chamber were lined with Perspex, behind which different types of wallpaper were hung. Working clockwise from the top-left chamber, the wallpapers in each chamber were as follows: black, spot (diameter: 15 mm; centre-to-centre distance: 25 mm), white, and check (29 mm × 29 mm squares). These wallpapered environments served as the four visual stimuli: A, B, C and D. Each chamber was equipped with a food well into which 45-mg food pellets could be delivered. A transparent plastic flap, 6 cm high × 5 cm wide, hinged along the top of the food well opening, guarded access to the food well. A movement of this flap of, approximately, 2 mm was automatically recorded as a single response or food well entry. The floors of the chambers were constructed from stainless steel rods (with diameters of 5 mm and mounted 15 mm apart); these rods could be electrified using a shock generator coupled with a shock scrambler (Campden Instruments Ltd., Loughborough, U.K., Model no: 521C and 521S, respectively). A speaker mounted on the ceiling of each operant chamber was used to present the auditory stimuli, X and Y. An aperture cut into the Perspex and aligned with the position of the speaker allowed for unimpeded delivery of sound. The two, 10-s auditory stimuli were a 10-Hz train of clicks (produced by one audio generator; Campden Instruments Ltd., Model no. 258) and a 2000-Hz constant tone (produced by a second and identical audio generator). These stimuli were presented at an intensity of, approximately, 75 dB (A weighting). A computer controlled the apparatus and recorded food well entries.

4.1.2.3 Procedure

Magazine training

Before the chambers were decorated, rats were trained to collect food pellets (Noyes Precision Pellets supplied by Sandown Chemicals Ltd, Hampton, England) from the food well over the course of two days. On the first day of training, the plastic flaps that guarded access to the food wells were fixed in a raised position to allow rats clear sight of, and easy access to, the food pellets. During the second day of training, the plastic flaps were lowered to their normal positions, and rats had to move the flaps to gain access to the food pellets. During both training sessions, 20 food pellets were delivered on a fixed-time 60-s schedule. The chambers were then decorated, and rats received 32 days of discrimination training.

Discrimination training

On each day of training, rats received one session of training with each of the four visual stimuli (A, B, C and D). Following completion of a session of training, rats were given experience with the next designated visual stimulus. For all rats, alternate days of training were conducted in the presence of auditory stimulus X and Y; in each session there were 10, 10-s presentations of either X or Y, and the interval between successive presentations within a session was 30-s. For half of the rats, presentations of X were immediately followed by the delivery of a single food pellet in A and B and were nonreinforced in C and D, and presentations of Y were reinforced in A and D, but not in B and C (see Table 6). For the remaining rats, the presentations of X were nonreinforced in A and B and reinforced in C and D, and those of Y were nonreinforced in A and D and reinforced in B and C. For all rats, the black and white visual stimuli served as A and C and the check and spot visual stimuli served as B and D. For half of the rats in the above subconditions, black served as A and white as C, and for the remainder this arrangement was reversed. For half of the rats in the subconditions created by the previous counterbalancing operations, check served as B and spot served as D, and for the remainder, the reverse was the case. The order in which rats received the four sessions within a day changed from one day to the next. Within an 8-day block of training, each visual stimulus was presented in the four possible positions within a day (1^{st}, 2^{nd}, 3^{rd}, 4^{th}), and experience with any one of the visual stimuli was equally likely to be immediately followed by or preceded by experience with any of the other three visual stimuli.

Revaluation and test

On the next two days, rats received aversive conditioning. On each day, rats received two sessions of training that were separated by a 2-hr interval. During one session, rats were placed in A where they received three 0.5-s, 0.5 mA electric shocks. Shocks were delivered at the rate of one every min. After approximately 30 s, rats were removed from A. In the other session, rats were simply placed in C for 3.5 min and were then removed. Within each of the sub-conditions created by the previous counterbalancing operations, for half of the rats, the orders in which A and C occurred were A, C (day 1) and C, A (day 2), and for the remainder the orders were C, A (day 1) and A, C (day 2). On both of the following two days, the behaviour of rats was video recorded during sessions in B and D that were separated by a 2-hr interval. On one day, rats received eight presentations of X that were separated by 10-s intertrial intervals, and on the other day, they received eight presentations of Y, again separated by 10-s intertrial intervals. This resulted in an overall session length of 2 min 40-s. Half of the rats received sessions with presentations of X on day 1 and Y on day 2, and the remainder received the reverse arrangement. Within the subconditions created by the previous counterbalancing operations, half of the rats received the sequence B, D on both days and the rest received the sequence D, B on both days.

Behavioural measures

Appetitive discrimination learning was assessed using the rate of food well entries (in responses per minute, rpm) during presentations of X and Y on the eight trial types. In fact, the eight trial types were separated into the simple discrimination (reinforced: AX & AY; nonreinforced: CX & CY) and the configural discrimination (reinforced: BX and DY; nonreinforced: BY and DX). Generalised fear, in the form of freezing behaviour, was assessed using a semi-automated scoring system reported in Grand and Honey (2008). Briefly, an observer (J.C.) watched the videotaped behaviour of rats from the test sessions and held down a mouse button when the rat moved, and released this button when the rat was stationary (i.e., freezing). Movement was defined as any behaviour with the exception of that necessary to maintain breathing. Each consecutive 2-s bin was scored as either containing a depression of the mouse button (i.e., the rat was active) or no depression of the mouse button (i.e., the rat was inactive or freezing). The trace of each rat's behaviour was then converted into the percentage of 2-s periods in which the rat was freezing. Data from 12 rats (18.75% of the data set) for Experiments 12 and 13 was second coded in order to assess inter-rater reliability. The second coder (R.C.H.) was blind with respect to the individual predictions for each rat. The inter-rater correlations (*r*) exceeded 0.95 in both Experiments 12 and 13, *p*s<.001.

4.1.3 Results

The results from the first stage of appetitive training are presented in 8-day blocks in Table 7. Inspection of Table 7 suggests that from the first block (involving 8 days) there was greater responding on the reinforced than on the nonreinforced trials and that this difference became more evident as training progressed for both the simple and configural discriminations. It is also apparent that the difference in responding between reinforced and nonreinforced trials for the simple discrimination was more marked than for the configural discrimination. ANOVA, with block (1-4), discrimination type (simple or configural), and reinforcement (+ or -) as factors, confirmed that there was an effect of discrimination type, $F(1, 15) = 46.13, p < .001$, an effect of reinforcement, $F(1, 15) = 151.73, p < .001$, but no effect of block, $F(3, 45) = 1.17, p > .05$. Furthermore, there was a significant interaction between discrimination type and reinforcement, $F(1, 15) = 13.84, p < .003$; no other interactions between factors were significant (largest $F(3, 45) = 2.19, p = .10$). Analysis of simple main effects revealed that whereas responding on the reinforced trials did not differ significantly between the two types of discrimination, $F(1, 15) = 1.61, p > .05$, responding on nonreinforced trials was significantly lower for the simple discrimination than for the configural discrimination, $F(1, 15) = 67.87, p < .001$. For both discrimination types, there was more responding on reinforced than on nonreinforced trials (smallest $F(1, 15) = 125.10, p < .001$).

Table 7. *Results from discrimination training in Experiments 12 and 13.*

Experiment 12				
Simple Discrimination				
Trial	Block 1	Block 2	Block 3	Block 4
+	4.20	4.07	4.12	4.14
-	0.84	0.65	0.59	0.35
Configural Discrimination				
Trial	Block 1	Block 2	Block 3	Block 4
+	4.28	4.14	4.44	4.36
-	1.88	1.48	1.32	1.21
Experiment 13				
Simple Discrimination				
Trial	Block 1	Block 2	Block 3	Block 4
+	4.79	3.81	4.12	4.67
-	1.08	0.73	0.73	0.70
Configural Discrimination				
Trial	Block 1	Block 2	Block 3	Block 4
+	5.17	4.16	4.25	4.57
-	2.29	1.62	1.44	1.25

Note. Mean rate of food well entries (in responses per minute, rpm) on reinforced (+) and nonreinforced (-) trials over each 8-day block, split by discrimination type (simple or configural).

Figure 26 shows the results of principal interest from Experiment 12: the amount of freezing elicited by B and D as a function of whether they were accompanied by auditory stimulus X or Y, pooled over the test periods. Inspection of this figure reveals that B elicited greater freezing than D when they were presented with auditory stimulus X, and that the reverse was true when they were presented with auditory stimulus Y. ANOVA, with visual stimulus (B or D) and auditory stimulus (X or Y) as factors, showed that while there were no main effects of visual or auditory stimulus (Fs<1), there was a significant interaction between these factors, $F(1, 15) = 23.89$, $p < .001$. Analysis of simple main effects confirmed that B elicited significantly greater freezing than D when they were presented with X, $F(1, 15) = 7.50$, $p < .02$, and that the reverse was true when they were presented with Y, $F(1, 15) = 11.19$, $p < .005$.

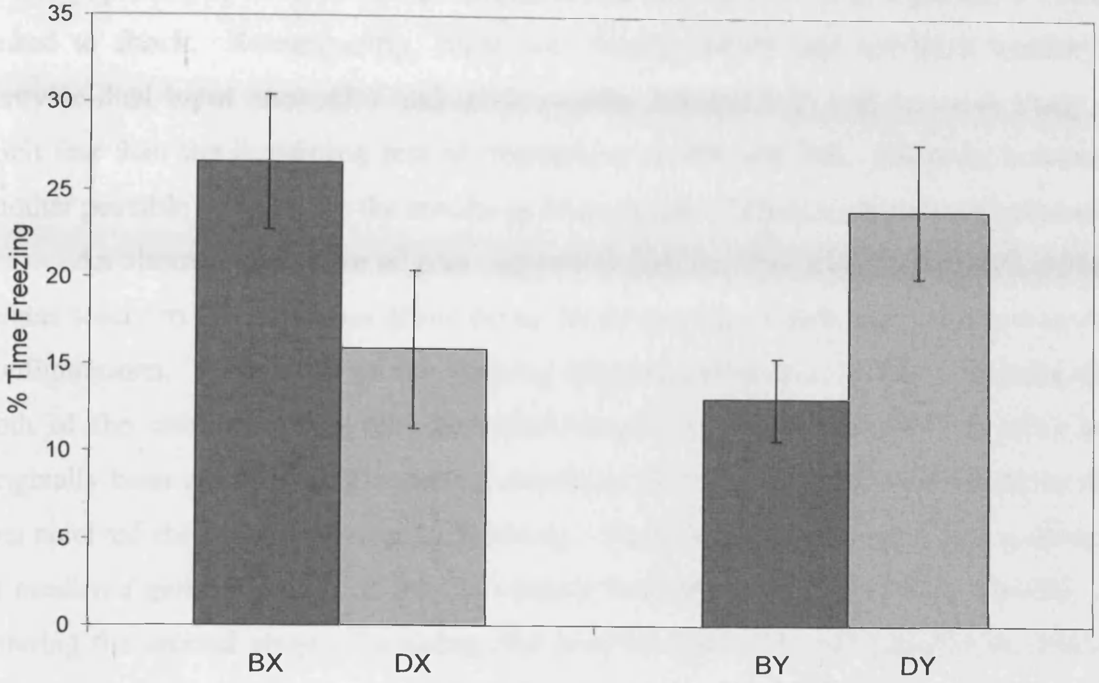

Figure 26. Results of Experiment 12: mean percentages of time freezing to visual stimuli B and D as a function of whether they were presented with auditory stimuli X or Y. Error bars indicate the standard error.

4.1.4 Discussion

Experiment 12 demonstrated that the generalisation of fear from A to the test stimuli (B and D) depended upon which auditory stimulus (X or Y) those test stimuli were accompanied by: generalisation of fear from A was more marked to BX than it was to DX, and it was more marked to DY than to BY. The sole way in which this pattern of results could have been generated is through the prior appetitive training. But, what feature of this prior training was critical?

According to the configural grouping account (detailed in Section 4), similar compounds (e.g., AX & BX) that are followed by the same outcome (e.g., food) become grouped – in the sense that they come to address a common configural unit (*ABX*). The appetitive training phase of Experiment 12, which involves eight trial types (e.g., AX→food, BX→food, CX→no food, DX→no food, AY→food, BY→no food, CY→no food, DY→food), should therefore result in pairs of similar compounds coming to address four configural units: *ABX* (for AX & BX), *CDX* (for CX & DX), *ADY* (for AY & DY), and *BCY* (for BY & CY). Once the network is configured in this way, pairing A with shock will activate *ABX* and /or *ADY*, and these units will be linked to shock.. Subsequently, those test configurations that are most similar to (provide dual input into) *ABX* and *ADY*, namely BX and DY, will be more likely to elicit fear than the remaining test configurations of BY and DX. There is, however, another possible account for the results of Experiment 12 that needs to be considered.

An alternative feature of prior appetitive training that might have been critical relates solely to the *outcomes* (food or no food) associated with the various stimulus configurations. Inspection of the training regime, outlined in Table 6, reveals that both of the configurations that provoked most freezing at test (BX & DY) had originally been paired with the same outcome as stimulus A (here food, for those rats that received the training shown in Table 6). These conditions might allow a process of *mediated generalisation* of fear to operate between A, BX, and DY. Briefly, on entering the second stage of training, the presentation of A will provoke activity in memory about the outcome with which it was paired during appetitive training (e.g., food). Under these conditions, A's pairing with shock might allow the associatively provoked memory of food to also become linked with shock. Therefore, when a representation of food is activated, for example, during presentations of BX and DY (but not of DX & BY), conditioned freezing may be generated. Although this analysis cannot explain many recent observations involving configural learning (e.g., Allman et al., 2004; Honey & Ward-Robinson, 2002; Honey & Watt, 1998, 1999), it has been applied to simple instances of the acquired equivalence and distinctiveness of cues (see, e.g., Honey & Hall, 1989). Experiment 13 was designed to discriminate between the two forms of analysis outlined above: one based on a process of configural grouping, and the other on simple mediated conditioning.

4.2 Experiment 13

4.2.1 Introduction

The design of Experiment 13 is summarised in the lower rows of Table 6. The first stage of training was identical to Experiment 12. However, during the second, revaluation stage, B was paired with shock and D was not, and in the final test stage, rats received test trials in which A and C were accompanied by X and Y. As can be seen from Table 6, B alone is uninformative about whether food or no food would be delivered during appetitive training. This means that when B is paired with shock, it is no more likely to activate a representation of food than of no food. However, if, for whatever reason, the associatively activated representation of food (or equally no food) did become linked to shock during the revaluation stage, then there would be no basis upon which to expect differential responding to A and C as a consequence of whether they are presented with X and Y: both AX and AY will provoke an associatively activated representation of one outcome (e.g., food), and both CX and CY will provoke an associatively activated representation of a different outcome (e.g., no food). This prediction contrasts with that made by the alternative configural grouping account. According to this account, B should activate the hidden units *ABX* and *CBY* during revaluation, resulting in these representations becoming linked to shock. Consequently, the test configurations that are most similar to *ABX* and *CBY*, namely AX and CY, respectively, should be more likely to elicit freezing than the remaining test compounds, AY and CX. Experiment 13 assessed these contrasting predictions.

4.2.2 Method

4.2.2.1 Subjects, apparatus and procedure

Sixteen naïve male hooded Lister rats (mean: 365g; range: 323-383g) that came from the same supplier and were maintained in the same way as those used in Experiment 12. The apparatus was that used in Experiment 12. All details of the experiment were the same as Experiment 12 with the following three exceptions: B was paired with footshock and D was not during the revaluation stage; AX, CX, AY and CY were presented during the test stage; and, the identities of the pairs of stimuli that served as A and C or B and D were exchanged in order to maintain the identities of the stimuli that were presented during revaluation (black and white) and test (check and spot).

4.2.3 Results

The results from the first stage of appetitive training, again divided according to discrimination type (simple or configural), are shown in Table 7. Inspection of this table suggests that, from the first block of training, there was more responding on the reinforced than on the nonreinforced trials, and again this difference increased as training progressed for both types of discrimination. Also, it is apparent that for the final two blocks of training, the difference in responding to the reinforced and nonreinforced trials was more pronounced for the simple discrimination than for the configural discrimination. ANOVA, with block (1-4), discrimination type (simple or configural), and reinforcement (+ or -) as factors, confirmed that there was a significant effect of block, *F*(3, 45) = 7.21, $p < .001$, discrimination type, *F*(1, 15) = 29.84, $p < .001$, and reinforcement, *F*(1, 15) = 201.14, $p < .001$. Also, significant interactions were revealed between block and discrimination type, *F*(3, 45) = 5.71, $p < .003$, and between discrimination type and reinforcement, *F*(1, 15) = 31.66, $p < .001$, but not between block and reinforcement, *F*(3, 45) = 2.56, $p = .07$. The three-way interaction between these factors was not significant, *F*<1. Analysis of simple main effects performed on the two significant interactions revealed that, for both types of discrimination, there was a significant reduction in responding on both rewarded and nonrewarded trials over block, smallest *F*(3, 13) = 3.74, $p < .05$. Also, for the first three blocks of training, the overall level of responding on the configural discrimination trials was significantly higher than the overall level of responding on the simple discrimination trials, smallest *F*(1, 15) = 14.83, $p < .003$; on the final block of training, this difference was not significant, *F*(1, 15) = 4.12, $p > .05$. As in Experiment 12, responding on reinforced trials did not differ significantly between the two types of discrimination, *F*(1, 15) = 1.94, $p > .05$, whereas responding on nonreinforced trials was significantly lower for the simple discrimination compared to the configural discrimination, *F*(1, 15) = 114.48, $p < .001$. For both types of discrimination, however, there was significantly more responding on reinforced trials than on nonreinforced trials (smallest *F*(1, 15) = 142.72, $p < .001$).

Figure 27 shows the test results from Experiment 13: the amount of freezing elicited by A and C as a function of whether they were presented with auditory stimulus X or Y, pooled over the test periods. Inspection of this figure reveals that A elicited greater freezing than C when these stimuli were presented with auditory stimulus X (i.e., AX elicited greater fear than CX), and that the reverse was true when they were presented with auditory stimulus Y (i.e., CY elicited greater fear than AY). ANOVA, with visual stimulus (A or C) and auditory stimulus (X or Y) as factors, found that there were no main effects of visual or auditory stimulus (*F*s<1), but that there was a significant interaction between these factors, *F*(1, 15) = 41.90, $p < .001$. Analysis of simple main effects confirmed that AX elicited greater freezing than CX, *F*(1, 15) = 11.13, $p < .006$, and CY elicited greater freezing than AY, *F*(1, 15) = 11.05, $p < .006$.

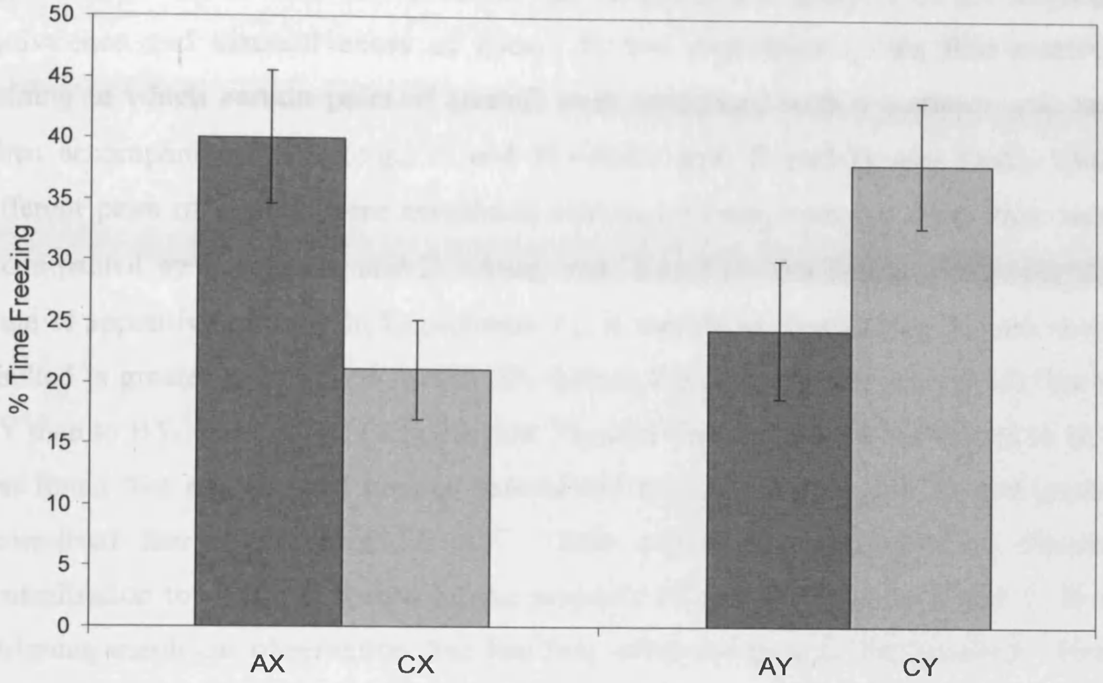

Figure 27. Results of Experiment 13: mean percentages of time freezing to visual stimuli B and D as a function of whether they were presented with auditory stimuli X or Y. Error bars indicate the standard error.

4.2.4 Discussion

The results of Experiment 13 confirm those of Experiment 12 in demonstrating a switch in similarity-based generalisation to two test stimuli (in this case A and C) that is dependent upon the stimulus (X or Y), or context, in which generalisation is assessed. They also allow us to discriminate between the two contrasting accounts proposed for the contextual modulation of stimulus generalisation shown in Experiment 12. Specifically, the results of Experiment 13 are inconsistent with an account based on simple mediated conditioning, and instead favour an account based on configural grouping. I will now consider in greater detail the implications of the results of Experiments 12 and 13.

4.3 General Discussion

The current experiments assessed the prediction that rats should be capable of forming groupings that allow for the cross-classification of a given set of stimuli (e.g., A, B, C and D), based upon the 'contextual' stimuli that accompany them (e.g., X or

Y). This prediction was derived from one connectionist analysis of the acquired equivalence and distinctiveness of cues. In two experiments, rats first received training in which certain pairs of stimuli were associated with a common outcome when accompanied by X (e.g., A and B→food, and, C and D→no food), while different pairs of stimuli were associated with a common outcome when they were accompanied by Y (e.g., A and D→food, and, B and C→no food). Following this stage of appetitive training in Experiment 12, it was found that pairing A with shock resulted in greater generalised fear to BX than to DX, and greater generalised fear to DY than to BY. Similarly in Experiment 13, after fear had been conditioned to B, it was found that rats showed greater generalised fear to AX than to CX, and greater generalised fear to CY than to AY. This contextual modulation of stimulus generalisation to A and C, based on the presence of auditory stimuli X and Y, is an intriguing empirical observation that has two, clear-cut general implications: First, these results provide further support for one connectionist analysis of learning and its application to the acquired equivalence and distinctiveness of cues (see Allman et al., 2004, Honey & Ward-Robinson, 2002). Second, they are clearly inconsistent with the suggestion that, unlike humans (Barsalou, 1982; Medin et al., 1993; Tversky & Gati, 1978), nonhuman animals are incapable of showing contextual modulation of similarity (cf. Chater & Heyes, 1994).

There are several possible ways in which the kind of connectionist approach described in the Introduction could be implemented (e.g., Allman et al., 2004; Gluck & Myers, 1993; Honey & Ward-Robinson, 2002), and there is independent support for some of these suggestions from procedures similar to those used in Experiments 12 and 13 (e.g., Honey & Ward-Robinson, 2001). However, it is worth noting that not all configural theories that have been implemented as connectionist networks are able to explain the results of Experiments 12 and 13. For example, Pearce's (1994) model supposes that each new pattern of stimulation (e.g., AX, BX, CX, DX, AY, BY, CY, DY) recruits a new configural (hidden) unit, and denies the possibility that there will be integration of configurations that are presented on different trials. For this model, therefore, the similarity between a given pair of configurations is fixed, being determined simply by the proportion of common elements that they share. However, one way in which similarity could be modified by experience in such a model is by allowing the outcome of a trial to be encoded as part of the configural representation of that trial (cf. Rescorla, 1991). This form of analysis is inconsistent with the results of other demonstrations of the acquired equivalence and distinctiveness of cues that use stimuli and procedures similar to those employed in Experiments 12 and 13 (see Honey & Ward-Robinson, 2001; see also, Delamater & Joseph, 2000; Hodder et al., 2003; Nakagawa, 1986; Urcuioli, Zentall & DeMarse, 1995; Zentall, Steirn, Sherburne, & Urcuioli, 1991). I therefore prefer the general suggestion, howsoever it is implemented, that similar patterns of stimulation that predict the same outcome come to address the same hidden unit, whereas otherwise equivalent patterns of stimulation that predict different outcomes come to address different hidden units (see Allman et al., 2004; Honey & Ward-Robinson, 2002).

As I have already mentioned, the results of Experiments 12 and 13 resonate with work in humans, where it is well established that similarity is not fixed in the manner prescribed by theories of stimulus generalisation in animals (e.g., Atkinson & Estes, 1963; McLaren & Mackintosh, 2000, 2002; Pearce, 1994). Instead, human judgements of similarity are highly flexible, being influenced by the context in which those judgements are made (e.g., Barsalou, 1982; Medin et al., 1993; Tversky & Gati, 1978). For example, a flashlight and a rope are only considered similar to one another when they are presented in the context *'taken on camping trips'*. I have presented a theoretical analysis of the contextual modulation of similarity in rats that appeals to relatively simple associative principles – albeit ones that are implemented within a three-layer connectionist network. Consequently, as well as establishing an important continuity in cognitive flexibility between humans and rats, these results also raise the intriguing possibility that the influence of context on human judgements of similarity may arise in an analogous fashion to that of rats. If human judgments of similarity are found to arise in an analogous fashion, then this would have important implications with respect to discussion about the role of the classifier in categorisation behaviour. Indeed, they would suggest that categorisation behaviour in all animals may stem from the same underlying, associative roots.

In Chapter 0, the idea was introduced that connectionist architectures may provide the best hope for providing a single model that can explain both human and nonhuman animal categorisation behaviour. The experiments presented in this chapter were developed from one connectionist architecture that affords a relatively high level of cognitive flexibility in nonhuman animals (Honey & Ward-Robinson, 2002), based on simple associative mechanisms (see also, Le Pelley, 2004). By finding this hypothesised continuity in cognitive flexibility between humans and rats, this lends hope to the aforementioned idea. With regards to the proposal of Chater and Heyes (1994) outlined at the beginning of this chapter, the shared influence of context on similarity across different species demonstrates that natural language is not a prerequisite for this form of complex, cognitively flexible behaviour. The results of Chapter 4, therefore, lend promise to the possibility that rats *do* have the cognitive requisites to engage in spontaneous categorisation. Moreover, this may well be in a manner that is qualitatively similar to incidental (spontaneous) categorisation behaviour in humans.

Chapter 5

General Discussion

5. Introduction

In this chapter I look to first summarise the main findings and implications of this thesis, and then look to the possible future research implications of my findings. Finally, I provide some general conclusions with respect to the findings presented in this thesis.

5.1 Summary and theoretical implications of the main findings

In this thesis, I investigated how stimulus similarity structure and the statistical properties of the environment influence categorisation behaviour in both humans and rats. In Chapter 1 of this thesis I conducted a review of the human literature on laboratory-based unsupervised categorisation and found that, for the most part, people show an overwhelming preference to engage in unidimensional unsupervised classification. This unidimensional unsupervised classification is odd, however, given that it does not conform with current understanding about the nature of people's everyday categories, which are built around a principle of family resemblance, and not 'definitions' (Rosch, 1975; Wittgenstein, 1953). While manipulations of stimulus format and experimental procedure (e.g., Milton & Wills, 2004; Milton et al., 2008), or the introduction of prior knowledge (e.g., Spalding & Murphy, 1996), have increased the prevalence of multidimensional (family resemblance) sorting, often, an overall preference for unidimensional classification remains. As noted in Chapter 1, however, one likely source of participants' unidimensional classification bias is the abstract similarity structure of the stimuli being classified. That is, I argued that the similarity-based relationships contained within a set of stimuli's abstract stimulus structure will likely command a strong influence over the issue of unidimensional versus multidimensional (family resemblance) classification. In Chapter 2 of this thesis, therefore, I investigated whether it was possible to predict unidimensional versus multidimensional unsupervised classification on the basis of abstract stimulus structure. In Chapter 3 of this thesis I looked to broaden the work of Chapter 2 by examining the general influence of stimulus similarity structure on whether a set of stimuli are spontaneously 'classified together' or spontaneously 'classified apart'.

More specifically, I introduced a new procedure to investigate some of the factors that might influence incidental unsupervised categorisation in both humans and nonhuman animals. Finally, given the findings of Chapter 3, in Chapter 4 of this thesis I examined the basic flexibility of rats' classification abilities when determined by the statistical properties of the environment.

5.1.1 Unidimensional versus multidimensional unsupervised categorisation

The dominance of unidimensional classification in studies of human unsupervised categorisation, identified in Chapter 1, has been the subject of much curiosity among categorisation researchers. Taking up this theme, Chapter 2 of this thesis investigated one likely factor in generating the laboratory-based unidimensional unsupervised classification bias; namely, the abstract similarity structure of the stimuli being classified. Specifically, I employed the simplicity model of unsupervised categorisation (Pothos & Chater, 2002) to predict when participants should prefer unidimensional classification and when they should prefer two-dimensional classification, on the basis of the abstract similarity structure of a set of stimuli.

In Experiment 1, participants' classifications were found to be more similar to the predicted 'suboptimal' ('less intuitive') category structure than the predicted 'optimal' ('more intuitive') category structure. That is, in the condition where simplicity predicted a unidimensional classification preference, participants' classifications were more similar to the predicted two-dimensional classification structure; in the condition where simplicity predicted a preference for two-dimensional classification, participants' classifications were more similar to the predicted unidimensional classification structure. However, for the stimulus structures employed in Experiment 1, the classification(s) predicted to be 'optimal' in each condition shared a superordinate-subordinate relationship with the classification(s) predicted to be 'suboptimal'. Consequently, it was not possible to determine whether participants' classification behaviour in Experiment 1 (which was opposite to the predictions of the simplicity model) represented a true preference, some unanticipated emergent dimension, or category subclustering. In Experiments 2 – 4, therefore, I sought to investigate these possibilities further. The results of Experiment 2 indicated that the results of Experiment 1 were unlikely to be the product of an unanticipated emergent dimension. Moreover, the results of

Experiments 3 and 4, which sought to reduce the likelihood of category subclustering, showed that participants' classification behaviour in Experiment 1 was robust to change. Critically, there was little evidence that participants initially classified in a manner that was consistent with the predictions of the simplicity model. One may argue, therefore, that this was indicative of a true preference, not necessarily for two-dimensional classification when simplicity predicted a preference for unidimensional classification, for example, but for category subclustering (see Gosselin & Schyns, 2001). While category subclustering is compatible with the simplicity model, it is still the case that participants' final classifications did not resemble most closely the classifications predicted to be 'optimal' by the simplicity model. Indeed, the simplicity model would never predict such category subclustering to be preferred by participants.

In the final experiment of Chapter 2, Experiment 5, two new stimulus structures were employed where the predicted 'optimal' classification(s) did not share a superordinate-subordinate relationship with the predicted 'suboptimal' classification(s). That is, the categorisation that represented classification along a single dimension of variation was as different as possible from the classification that took into account both dimensions of variation together. The results of Experiment 5 were found to support the predictions of the simplicity model. That is, where simplicity predicted a preference for unidimensional classification, participants' classifications were most similar to the predicted 'optimal' unidimensional classifications. Where simplicity predicted a preference for two-dimensional classification, participants' classifications were most similar to the predicted 'optimal' two-dimensional classification.

The findings from Chapter 2 of this thesis have a number of important theoretical implications for our understanding of human unsupervised categorisation. First, the experiments of Chapter 2 highlight the important influence of abstract similarity structure on human unsupervised classification. This is exemplified by the results of Experiment 5 in which I documented the first empirical demonstration showing a two-dimensional bias in unsupervised classification, on the basis of the abstract stimulus structure. Moreover, the results of Experiment 5 support the assumptions of Rosch (1975) that i) people engage in category construction by considering the similarity among a set of stimuli, and ii) that 'good' categories are those that maximise within-category similarity and minimise between-category similarity. Second, the results of Experiments 1 – 4 document that human categorisation does not always fit these assumptions. As shown, when there is meaningful substructure contained within the presumed 'optimal' classification (i.e., the classification that maximises within-category similarity and minimises between-category similarity), then participants may likely engage in category subclustering. This means that participants' final classifications will often reflect the subordinate level, presumed 'suboptimal' classification. Indeed, the results of Experiment 4, and to a lesser extent Experiment 3, show that this preference to generate classification hierarchies (see Gosselin & Schyns, 2001) is rather robust. Third, therefore, the findings of Chapter 2 clearly question the validity of the simplicity model. That is, while the simplicity model's predictions appear accurate under certain sets of conditions, under others, it will never correctly predict participants' classification behaviour (although, seeking category subclusters is, at least, compatible with the model). Overall, therefore, the findings of Chapter 2 appear to suggest that human category constructions is a product of an interaction between the processing biases of the classifier and the similarity structure of the stimuli (Ahn & Medin, 1990; see Love et al., 2004).

Finally, with respect to the overwhelming prevalence of unidimensional classification in the human unsupervised categorisation literature, the findings of Chapter 2 strongly suggest that this has arisen partly because of a lack in understanding of the biases that exist within the abstract similarity structure of a set of stimuli. Specifically, Chapter 2 showed that, as for SUSTAIN (Love et al., 2004), the simplicity model also predicts a unidimensional classification preference for the widely employed stimulus structure of Medin et al. (1987; see Figure 1, Chapter 1). Consequently, while Medin et al. (1987) assumed that this stimulus structure should promote family resemblance sorting – because people would construct categories around the category prototypes – actually, the structure was biasing people towards unidimensional classification. This confusion has therefore fostered a sense that people are acting oddly in many laboratory-based unsupervised categorisation studies, when in fact they are classifying the stimuli in the most intuitive way, on the basis of abstract similarity structure.

5.1.2 Within-category similarity structure and incidental categorisation

In Chapter 3 of this thesis, I investigated incidental categorisation in humans and rats. The reasons for taking this comparative approach were two-fold: first, it allowed for an examination of whether the mechanisms that underlie incidental categorisation in humans might also underlie incidental categorisation in rats. Second, therefore, it allowed for an assessment of the role of the classifier in spontaneous categorisation.

The findings of Experiment 6 appeared to support the view that humans are sensitive to a surprise-driven category invention mechanism in incidental categorisation (Clapper & Bower, 1994, 2002). That is, participants preexposed to a stimulus similarity structure that contained three highly similar stimuli (e.g., A, B, and C) and one distinct stimulus (e.g., F) showed a reduced amount of later property generalisation between stimuli A and F than participants preexposed only to stimuli A and F. This finding is consistent with the assumption that participants in the former group 'classified apart' stimuli A and F, and one presumes 'classified together' stimuli A, B, and C. Moreover, Experiment 7 confirmed that this result was the product of the distinct perceptual discontinuity created by exposure to, for example, stimuli A, B, C, and F, and not the result of the temporal discontinuity that existed in Experiment 6 between exposure to stimulus C and exposure to stimulus F, for example. Although temporal factors were not critical in determining the incidental categorisation behaviour of participants in Experiments 6 and 7, Experiment 8 demonstrated that temporal factors play an important role in human incidental categorisation. Specifically, Experiment 8 showed that increasing the temporal contiguity between the presentation of two similar, but distinct stimuli (e.g., A and F) increases their subsequent perceived similarity to each other.

Following this work in humans, Experiments 9 – 11 examined incidental categorisation in rats. Overall, the pattern of results observed in rats was qualitatively different from that observed in humans. Specifically, in Experiment 9, an effect of perceptual learning was observed. That is, rats that received preexposure to all of the four tone stimuli employed (i.e., A, B, C and D) showed a reduced amount of later generalisation between stimuli A and D than those rats that were preexposed only to stimuli A and D, and to A, B, and D, for example. In Experiment 10, the temporal properties of stimulus preexposure were found to command a strong influence over rats' later generalisation behaviour. Specifically, by massing stimulus exposure, rats that received exposure to all four tone stimuli (in condition Sys_trans) came to show somewhat greater levels of generalisation between stimuli A and D than rats that received exposure only to stimuli A and D. This result was confirmed in Experiment 11. By combining Experiments 10 and 11, it was possible to conclude that rats in condition Sys_trans (i.e., those that received preexposure to stimuli A, B, C and D) subsequently showed significantly more generalisation between stimuli A and D relative to rats that only received preexposure to stimuli A and D.

The implications, both theoretical and practical, of Chapter 3 are broad. First, a new experimental procedure was introduced that allowed for a formally equivalent, comparative assessment of incidental categorisation in humans and rats. Second, although the patterns of results were qualitatively different, the findings of Chapter 3 showed that the stimulus similarity structure of an exposed set of stimuli commands an important influence over later generalisation behaviour in both humans and rats. Specifically, the findings of Experiments 6 and 7 support the view that humans engage in incidental category formation on the basis of stimulus similarity structure. Moreover, they appear consistent with the proposal of Rosch (1975) that people prefer to form categories that maximise within-category similarity and minimise between-category similarity. Experiments 6 and 7 lend strong support to the proposal of Clapper and Bower (1994, 2002) that human spontaneous categorisation is guided by a surprise-driven category invention mechanism (see also Love et al., 2004). Interestingly, the patterns of results from Experiments 6 – 8 do not support the proposal that transformational knowledge encourages stimuli to be 'classified together' (cf. Zaki & Homa, 1999).

One of the most important findings of Chapter 3 was that, using a formally equivalent experimental design to Experiment 6, rats showed little evidence of meaningful, human-like incidental categorisation. Rather, simpler associative mechanisms could readily explain the rat results of Experiments 9 – 11. Of course, this is not to say that other nonhuman animals would not show incidental categorisation behaviour that is consistent with the findings from humans. Overall, however, the findings of Chapter 3 suggest an important role for the classifier in incidental categorisation.

5.1.3 Cross-classification in rats

In Chapter 4 of this thesis, I investigated whether rats exhibited another important aspect of human categorisation; namely, stimulus cross-classification. To recapitulate, some authors have argued that, on the basis of simple associative processes, nonhuman animals are incapable of meaningful, human-like categorisation (see Chater & Heyes, 1994). In particular, whereas human categorisation is effortlessly flexible, nonhuman animal categorisation will be inflexible, due to the nature of association formation. If one accepts this argument, it is hardly surprising that rats did not show incidental classification behaviour that was consistent with that of humans in Chapter 3. However, recent experimental results (see Honey & Watt, 1998, 1999), and the connectionist architectures born from this work, have challenged the arguments of Chater and Heyes (1994). In Chapter 4 of this thesis, therefore, I examined one prediction from the connectionist architecture outlined by Honey and Ward-Robinson (2002). Specifically, this architecture predicts that simple associative processes should afford flexible forms of categorisation behaviour, such as stimulus cross-classification. Over the two experiments detailed in Chapter 4, I showed that rats do have the cognitive requisites to engage in stimulus cross-classification, by demonstrating that rats' perceptions of similarity are context-dependent.

The results of Chapter 4, therefore, supported the connectionist analysis of learning outlined by Honey and Ward-Robinson (2002), demonstrating that relatively simple associative principles can bring about complex, cognitively flexible forms of classification behaviour. This has important implications with respect to the plausibility of spontaneous categorization in nonhuman animals. Specifically, the findings of Chapter 4 are, at least, suggestive of the possibility that spontaneous categorisation is not beyond the cognitive capacities of rats. Moreover, Experiments 12 and 13 of Chapter 4 have two further theoretical implications: First, they show that natural language is not a prerequisite for complex, cognitively flexible cognition. Second, they raise the question of whether the context-dependent nature of similarity in humans arises through similar associative processes.

5.2 Suggestions for future research

The experiments detailed in this thesis raise many interesting questions for future research. In this section, therefore, I propose a number of avenues for future research that would help to explore more fully some of the theoretical implications raised from this thesis.

5.2.1 Unidimensional versus multidimensional unsupervised categorisation

In a general sense, the results of Chapter 2 of this thesis demonstrate that further empirical investigation of the influence of abstract stimulus structure on unsupervised categorisation is an important topic for future research. A number of more specific suggestions for future research are apparent from the experiments of Chapter 2, however, and these are discussed further below.

One particularly interesting topic for future research concerns the issue of category subclustering (or subordinate level categorisation) in human unsupervised categorisation; this issue arose in Experiments 1 – 4 of Chapter 2. One immediate question is as follows: when employing stimulus structures where the predicted 'optimal' classification(s) shares a superordinate-subordinate relationship with the predicted 'suboptimal' classification(s) (at least, according to the simplicity model), is it the case that participants' final classifications will *always* be more similar to the 'suboptimal' (subordinate level) classification(s)? Moreover, following up the work conducted in Experiments 3 and 4 of Chapter 2, what factors influence the occurrence of category subclustering in unsupervised categorisation? Milton et al. (2008), for example, have shown that taxing working memory influences participants' classification behaviour. If one were to impose a strict time constraint on classification *and* tax working memory at the same time, would this be sufficient to eradicate any influence of category subclustering, and therefore reverse the results of Experiments 1 – 4?

When the issue of category subclustering is negated, the simplicity model of unsupervised categorisation accurately predicts when participants will prefer unidimensional classification, and when they will prefer multidimensional classification (see Experiment 5, Chapter 2). A number of obvious follow-ups to this finding present themselves: First, it is important to establish if the simplicity model makes accurate predictions with respect to binary dimensioned, as opposed to continuous dimensioned, stimuli. As documented in Chapter 2, the simplicity model accurately predicted a bias for unidimensional unsupervised classification for the four dimensional, binary-valued stimulus structure employed by Medin et al. (1987; see Figure 1, Chapter 1). However, it is not known whether participants would preferentially engage in family resemblance sorting when presented with a binary-valued stimulus structure for which simplicity predicts a preference for multidimensional unsupervised classification. The fact that so many unsupervised categorisation studies have employed binary-valued stimuli makes research of this nature particularly interesting, although not the most naturalistic. Second, it would be interesting to extend this work into stimuli with more than two dimensions. Obviously natural stimuli are composed of many different dimensions of variation; consequently, to study unsupervised categorisation in a naturalistic way in the laboratory, those stimuli employed should similarly have many dimensions of variation. Unfortunately, modelling work becomes particularly complex when employing stimuli with a great number of dimensions of variation. In future research, though, modelling work involving three dimensions of variation would be an obvious next step.

Third, focusing on the stimulus structures employed in Experiment 5 of Chapter 2, it would be interesting to see whether participants' classification preferences were observable in a supervised learning task. That is, in the condition where simplicity predicted a preference for unidimensional classification, is it the case that people learn the predicted classifications along either just dimension *x* or just dimension *y* more quickly than the predicted classification that takes into account both dimensions of variation together? Similarly, in the condition where simplicity predicted a preference for two-dimensional classification, is it the case that people learn the predicted classification that takes into account both dimensions of variation together more quickly than the predicted classification along either just dimension *x* or just dimension *y*? If people do show faster learning of the classification(s) that are predicted to be 'optimal' in each condition, then it would be a simple matter to employ the same learning task in, for example, pigeons or nonhuman primates. This interesting research would therefore allow for an assessment of unidimensional versus multidimensional classification in nonhuman animals. Moreover, it would allow for a direct comparative contrast of the classification biases that arise from abstract stimulus structure in human and nonhuman animals.

Finally, in future research I would like to introduce a time constraint on people's unsupervised classification of the stimuli derived from the two stimulus structures employed in Experiment 5. To recapitulate, time pressure has been associated both with an increase in unidimensional unsupervised classification (e.g., Milton et al., 2008), and an increase in unsupervised classification based on a principle of family resemblance (e.g., Smith & Kemler Nelson, 1984). In Experiment 5 of Chapter 2, one situation was established that promoted a unidimensional classification bias, and a second situation was established that promoted a two-dimensional classification bias. Consequently, these conditions provide a perfect situation through which to assess whether a time constraint should be associated with a unique increase in unidimensional sorting, or a unique increase in family resemblance sorting. If it should be associated with an increase in unidimensional unsupervised classification, then one should expect an enhancement in unidimensional classification in the condition where such a bias is predicted, and a reduction in two-dimensional classification in the condition where such a bias is predicted. In contrast, if imposing a time constraint should be associated with an increase in family resemblance sorting, then one should expect a reduction in unidimensional classification in the condition where such a bias is predicted, and an increase in two-dimensional classification in the condition where such a bias is predicted. Equally plausible, however, is the possibility that a time constraint will simply lead to an enhancement of the results of Experiment 5, due to people being forced to only classify on the basis of the classification(s) that they perceive to be more intuitive, or 'optimal'. This research would have wide implications for our understanding of the basis of unsupervised category construction, as it would allow for a clear assessment of which classification strategy should be regarded as the 'primitive' of human unsupervised categorisation.

5.2.2 Within-category similarity structure and incidental categorisation

A number of important theoretical implications of the findings from Chapter 3 were considered in Section 5.1.2. With these implications in mind, a number of follow-up studies present themselves for future research.

First, the findings of Experiments 6 – 8 in humans need to be replicated with a different stimulus set. This is important to establish that the findings were not a product of some particular quality of the stimulus set employed in these experiments, and to also confirm the robustness of the findings. In particular, it would be interesting to employ stimuli that were entirely arbitrary, meaning that there was absolutely no associated prior knowledge with the stimuli being classified (e.g., Grand et al., 2007). Second, it would be interesting to extend the stimulus set used to try to increase the size of the effects found. Based on the theorising of Clapper and Bower (1994, 2002), for example, the surprise-driven category invention mechanism should be more effective the more information people have about the 'norms' that establish membership in Category A, and the norms that establish membership in Category B. Consequently, it would be interesting to replicate Experiment 6 employing a stimulus set that contains renderings of morph stimuli at 1%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. Given such a stimulus set, one would then be able to expose participants in the Surprise condition to the 1%, 10%, 20%, 30%, 40%, 80%, and 100% morph renderings, for example. This should enhance the perceived perceptual discontinuity between the 1%, 10%, 20%, 30% and 40% stimuli and the 80% and 100% stimuli, meaning that participants in the Surprise condition should be more definite in their belief that the 1% and 100% stimuli should be classified in separate categories. If true, then later generalisation between the 1% and 100% stimuli should be reduced to a greater extent than that found in Experiment 6.

In future research, I would also like to introduce a fifth condition that explores participants' perceptions of the 1% and 100% stimuli after viewing morph animations of the 1% stimuli morphing into the 100% stimuli. To recapitulate, in Experiments 6 – 8 of Chapter 3, no effect of transformational knowledge was found. It would be of particular interest, therefore, if viewing the complete morph animation (i.e., the 1% stimuli morphing into the 100% stimuli) did influence later generalisation between these stimuli, relative to a baseline condition. Indeed, there is good reason to think that it might: recent research has shown that exposure to short animations of one object morphing into another object does influence participants' perceptions of stimulus similarity (Hahn et al., 2009; see also, Hockema, Blair, & Goldstone, 2005).

The findings of Chapter 3 also raise the interesting question of whether incidental categorisation, guided by a surprise-driven category invention mechanism, is prevalent throughout human development. Due to the flexibility of the experimental procedure introduced to study incidental categorisation in Chapter 3, this should be a relatively simple question to test. For example, following stimulus preexposure in older children, one could establish a small startle response (e.g., loud noise) to, for example, the 1% stimulus. Subsequently, one could monitor the child's response during presentation of the 100% stimulus. The prediction would be that the greater the perceived similarity of the 1% and 100% stimuli, the greater the startle response will be to the 100% stimulus, in this case. The experimental procedure could also be adapted to the study of incidental classification in infants and neonates. Specifically, following stimulus preexposure, one could then present, for example, the 1% stimulus for a period of time until the infant/ neonate loses interest in the stimulus. Subsequently, one could monitor the infant's/ neonate's looking time to the 100% stimulus, in this case. Critically, novelty has been found to be closely related to an increase in looking time (e.g., Behl-Chadha, 1996; Murai et al., 2004). The prediction here would be, therefore, that the more similar the 1% and 100% stimuli are perceived to be, the less novel the 100% stimulus will appear, which will reduce infant/ neonate looking time to this stimulus. This research would be of particular interest to the question of whether humans engage in category construction first, and then later apply language labels to these categories, or whether category construction is based on previously learned language labels (see Nelson, 1974).

The results of Experiments 9 – 11 in rats raise a number of interesting questions for future research. First, do other nonhuman animals engage in incidental categorisation in a manner that is consistent with the mechanisms that appear to underlie incidental categorisation in humans? Critically, the experimental procedure developed in Chapter 3 can be employed to assess incidental categorisation in many different species of nonhuman animal. To reiterate from Chapter 3, important differences existed between the assessment of incidental categorisation in humans (Experiments 6 – 8) and the assessment of incidental categorisation rats (Experiments 9 – 11). Most obvious of these differences is that while humans received exposure to visual stimuli, rats received exposure to auditory stimuli. In future research, therefore, I am keen to employ the experimental procedure introduced in Chapter 3 to examine incidental categorisation in pigeons and nonhuman primates. By focusing on these animals, it would be possible to expose exactly the same visual stimuli that were exposed to the human participants in Experiments 6 – 8. Consequently, a more direct comparative assessment of incidental categorisation in human and nonhuman animals would be possible. Second, the findings of Experiments 10 and 11 showed some evidence that transformational information increased the perceived similarity of stimuli A and D in rats, relative to a baseline condition. Specifically, those rats preexposed to stimuli A, B, C and D showed a greater amount of subsequent generalisation between stimuli A and D than rats preexposed only to stimuli A and D. This finding was brought about by massing stimulus preexposure to a greater extent than in Experiment 9. In future research, therefore, I would like to mass stimulus preexposure further by exposing rats to all of the stimuli they are scheduled to receive within a single session. That is, exposure to stimulus B would occur immediately after exposure to stimulus A, and so on. Moreover, future research could look to play a dynamic auditory stimulus to rats, where stimulus A is heard to transform into stimulus D. To the best of my knowledge, nonhuman animals have not been preexposed to transformational stimuli such as this, which makes research of this type particularly interesting.

Finally, in future research I hope to investigate incidental categorisation further using a within-participants version of the incidental classification task introduced in Chapter 3, which I have developed. A within-participant design would be particularly useful when assessing incidental categorisation in nonhuman animals, as having sufficient experimental power is always an issue.

5.2.3 Cross-classification in rats

The experimental results of Chapter 4 raise the intriguing possibility that the influence of context on human judgements of similarity may arise on the basis of simple associative processes (i.e., in an analogous way to that of rats). One possible way to test this would be to adapt the experimental task employed by Grand et al. (2007). For example, one could have participants initially learn a discrimination in which different 'spacebugs' are killed by different insecticide sprays. That is, while in context X, a red insecticide spray kills spacebug stimuli A and B and a blue insecticide spray kills spacebug stimuli C and D, in context Y, the red insecticide spray kills spacebug stimuli A and D, and the blue insecticide spray kills spacebug stimuli B and C. After learning this discrimination, one could then teach participants that only a new yellow insecticide spray will now kill spacebug stimulus B, while a new green insecticide spray is required to kill spacebug stimulus D. Subsequently, generalisation of the use of the yellow and green insecticide sprays can be assessed to spacebug stimuli A and C within each context (i.e., X and Y). The prediction here would be that, in context X, participants should come to use the yellow insecticide spray to kill spacebug stimulus A and the green insecticide spray to kill spacebug stimulus C. In contrast, in context Y, participants should come to use the yellow insecticide spray to kill spacebug stimulus C and the green insecticide spray to kill spacebug stimulus A.

The findings of Experiments 12 and 13 of Chapter 4 supported the predictions of the connectionist analysis outlined by Honey and Ward-Robinson (2002). In future research, I also hope to interrogate this connectionist architecture further to see whether other unique predictions are made about the flexibility of nonhuman animal behaviour, based on simple associative principles.

5.3 Conclusions

This thesis investigated how stimulus similarity structure and the statistical properties of the environment influence certain categorisation behaviour in humans and rats. With respect to humans, the experimental work presented in this thesis has shown that stimulus similarity structure commands an important influence over our unsupervised categorisation behaviour. Indeed, stimulus similarity structure is likely to be a key determinant of the overwhelming bias for unidimensional unsupervised classification found in the laboratory. This influence of stimulus similarity structure on human unsupervised categorisation makes sense; to quote Anderson, "psychologists must understand human behaviour by assuming it is adapted to the environment" (1991, p. 409). That is, Anderson (1991) argues that the human mind is adapted to pick up on perceived regularity within the environment, and that this perceived regularity will be utilised by humans. However, the findings of Chapters 2 and 3 of this thesis also show that human categorisation is a complex phenomenon, which is influenced by many different factors. Consequently, while categorisation based simply on a principle of maximising within-category similarity and minimising between-category similarity is appealing, it can only take one so far. The complexity of human categorisation is where flexible models, such as SUSTAIN (Love et al., 2004), come to the fore, while more inflexible models, such as the simplicity model (Pothos & Chater, 2002), show their limitations. With respect to the simplicity model of unsupervised categorisation, ultimately I believe that the combinatorics involved in the model will let it down. The amount of processing power that would be required to deal with real world categorisation, in which one sees thousands of dimensionally complex stimuli, is vast. Perhaps, however, this is where 'knowledge' factors may play a role in the model. Specifically, prior knowledge may be able to provide further constraints on categorisation, which may radically reduce the amount of processing power required to deal with natural categorisation. To reiterate though, the interaction between general knowledge and unsupervised categorisation is an incredibly complex process (see, e.g., Heit, 1997; Malt & Sloman, 2007). To accommodate general knowledge factors, therefore, the simplicity formalism would need considerable revision.

With respect to rats, the experimental work presented in this thesis has shown that they are sensitive to stimulus similarity structure, but in a qualitatively different way to humans. On the one hand, therefore, little evidence was found to support the idea that rats engage in spontaneous categorisation in a manner that is similar to humans, or even at all. On the other hand, however, the findings of Chapter 4 show that rats can show complex, cognitively flexible form of classification behaviour. To my mind, I believe that the findings of Chapter 4 are at least suggestive of the possibility that rats (and most likely other nonhuman animals) do have the cognitive requisites to engage in some rudimentary form of spontaneous categorisation; the problem is how to reveal this. With respect to this problem, I hope that the experimental procedure introduced in Chapter 3 will be at the forefront of future investigations of spontaneous categorisation behaviour in nonhuman animals.

To conclude, as highlighted at the beginning of this thesis, the connectionist analysis of human and nonhuman behaviour is revealing interesting new avenues for future research all the time. I believe that such analysis will be critical if we ever hope to have a unified theory and model of human and nonhuman animal categorisation. Of course, this is still a long way off, but with targeted comparative research, it may just happen.

References

Ahn, W.K. (1990). Effects of background knowledge on family resemblance sorting. *Proceedings of the 12th Annual Conference of the Cognitive Science Society* (pp. 149-156). Hillsdale, NJ: Erlbaum.

Ahn, W.K. (1991). Effects of background knowledge on family resemblance sorting: Part II. *Proceedings of the 13th Annual Conference of the Cognitive Science Society* (pp. 203-208). Hillsdale, NJ: Erlbaum.

Ahn, W.K., & Medin, D.L. (1992). A two-stage model of category construction. *Cognitive Science, 16,* 81-121.

Atkinson, R.C., & Estes, W.K. (1963). Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology: Vol. 2.* New York: Wiley.

Allman, M., Ward-Robinson, J., & Honey, R.C. (2004). Associative change in the representations acquired during conditional discriminations: Further analysis of the nature of conditional learning. *Journal of Experimental Psychology: Animal Behavior Processes, 30,* 118-128.

Anderson, J.R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J.R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.

Armstrong, S.L., Gleitman, L.R., & Gleitman, H. (1983). What some concepts might not be. *Cognition, 13,* 263-308.

Ashby, F.G., Queller, S., & Berretty, P.M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61,* 1178-1199.

Astley, S.L., & Wasserman, E.A. (1992). Categorical discrimination and generalization in pigeons: All negative stimuli are not created equal. *Journal of Experimental Psychology: Animal Behavior Processes, 18,* 193-207.

Aydin, A., Pearce, J.M. (1994). Prototype effects in categorisation by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 20,* 264-277.

Banfield, C.F., & Bassill, S. (1977). A transfer algorithm for non-hierarchical classification. *Applied Statistics, 26,* 206-210.

Barrett, P.T., Petrides, K.V., Eysenck, S.B.G., Eysenck, H.J. (1998). The Eysenck personality questionnaire: An examination of the factorial similarity of P, E, N, and L across 34 countries. *Personality & Individual Differences, 25,* 805-819.

Barsalou, L.W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition, 10,* 82-93.

Bateson, P.P. (1973). Internal influences on early learning in birds. In R.A. Hinde & J.S. Hinde (Eds.), *Constraints on learning* (pp. 101-116). London: Academic Press.

Bateson, P.P., & Chantrey, D.F. (1972). Retardation of discrimination learning in monkeys and chicks previously exposed to both stimuli. *Nature, 237,* 173-174.

Behl-Chadha, G. (1996). Basic-level and superordinate-like categorical representations in early infancy. *Cognition, 60,* 105-141.

Bennett, C.H., & Mackintosh, N.J. (1999). Comparison and contrast as a mechanism of perceptual learning? *Quarterly Journal of Experimental Psychology, 52B,* 253-272.

Berlin, B., Breedlove, D.E. & Raven, P.H. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist, 75,* 214-242.

Billman, D. (1989). Systems of correlations in rule and category learning: Use of structured input in learning syntactic categories. *Language and Cognitive Processes, 4,* 127-155.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Human Learning & Memory, 22,* 458-475.

Bornstein, M.H. (1987). Perceptual categories in vision and audition. In S. Harnad (Ed.), *Categorical Perception: The Groundwork of Cognition* (pp. 535-565). New York: Cambridge University Press.

Brown, D.A., & Boysen, S.T. (2000). Spontaneous discrimination of natural stimuli by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology, 114,* 393-400.

Bruner, J.S., Goodnow, J., & Austin, G.A. (1956). *A study of thinking.* New York: Wiley.

Calder, A.J., Young, A.W., Perrett, D.I., Etcoff, N.L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3,* 81-117.

Cerella, J. (1979). Visual classes and natural categories in the pigeon. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 68-77.

Chantrey, D.F. (1972). Enhancement and retardation of discrimination learning in chicks after exposure to the discriminanda. *Journal of Comparative and Physiological Psychology, 81,* 256-261.

Chantrey, D.F. (1974). Stimulus preexposure and discrimination learning by domestic chicks: Effect of varying interstimulus time. *Journal of Comparative and Physiological Psychology, 87,* 517-525.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology, 52A,* 273-302.

Chater, N., & Heyes, C. (1994). Animal Concepts: Content and Discontent. *Mind & Language, 9,* 209-246.

Cheeseman, P., & Stutz, J. (1995). Bayesian classification (AutoClass): Theory and results. In M.F. Usama, P.S. Gregory, S. Padhraic, & U. Ramasamy (Eds.), *Advances in knowledge discovery and data mining.* Menlo Park: The AAAI Press.

Cheng, K., Shettleworth, S.J., Huttenlocher, J., & Rieser, J.J. (2007). Bayesian integration of spatial information. *Psychological Bulletin, 133,* 625-637.

Clapper, J.P. (2007). Prior knowledge and correlational structure in unsupervised learning. *Canadian Journal of Experimental Psychology, 61,* 109-127.

Clapper, J.P, & Bower, G.H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 443-460.

Clapper, J.P., & Bower, G.H. (2002). Adaptive categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 908-923.

Close, J., Hahn, U., Hodgetts, C.J., Pothos, E.M. (2009). Rules versus similarity in concept learning. In D. Mareschal, P.C. Quinn, & S.E.G. Lea (Eds.), *The making of human concepts* (pp. 29-52). Oxford University Press.

Compton, B.J., & Logan, G.D. (1993). Evaluating a computational model of perceptual grouping. *Perception & Psychophysics, 53,* 403-421.

Compton, B.J., & Logan, G.D. (1999). Judgments of perceptual groups: Reliability and sensitivity to stimulus transformation. *Perception & Psychophysics, 61,* 1320-1335.

Corter, J., & Gluck, M. (1992). Explaining basic categories: feature predictability and information. *Psychological Bulletin, 111,* 291-303.

D'Amato, M.R., & Van Sant, P. (1988). The person concept in monkeys (*Cebus apella*). *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 43-56.

Delamater, A.R., & Joseph, P. (2000). Common coding in symbolic matching tasks in humans: Training with a common consequence or antecedent. *Quarterly Journal of Experimental Psychology, 53B,* 255-273.

Dougherty, J.W.D. (1978). Salience and relativity in classification. *American Ethnologist, 5,* 66-80.

Dwyer, D.M., Hodder, K.I., & Honey, R.C. (2004). Perceptual learning in humans: Roles of pre-exposure schedule, feedback, and discrimination assay. *Quarterly Journal of Experimental Psychology, 57B,* 245-259.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Elman, J.L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7,* 195-225.

Fagot, J., Wasserman, E.A., & Young, M.E. (2001). Discriminating the relation between relations: The role of entropy in abstract conceptualization by baboons and humans. *Journal of Experimental Psychology: Animal Behavior Processes, 27,* 316-328.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology, 41,* 145-170.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407,* 630-633.

Field, A.P. (2009). *Discovering statistics using SPSS: and sex and drugs and rock 'n' roll* (3^{rd} ed.). London: Sage.

Finke, R.A., Freyd, J.J., & Shyi, G.C.W. (1986). Implied velocity and acceleration induce transformations of visual memory. *Journal of Experimental Psychology: General, 115,* 175-188.

Fisher, D. (1987). Knowledge acquisition *via* incremental conceptual clustering. *Machine Learning, 2,* 139-172.

Fisher, D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research, 4,* 147-179.

Fisher, D., & Langley, P. (1990). The structure and formation of natural categories. In B. Gordon (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 241-284). San Diego, CA: Academic Press.

Fodor, J.A. (1983). The modularity of mind: an essay on faculty psychology. Cambridge, MA: MIT Press.

Fodor, J.A., Garrett, M.F., Walker, E.C.T., & Parkes, C.H. (1980). Against definitions. *Cognition, 8,* 263-367.

Frege, G. (1970). On sense and reference, translated by M. Black. In P. Geach & M. Black (Eds.), *Philosophical Writings of Gottlob Frege.* Oxford, Basil: Blackwell. Original publication,1892.

Freyd, J.J., & Finke, R.A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 126-132.

Fried, L.S., & Holyoak, K.J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10,* 234-257.

Garner, W.R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Gibson, E.J. (1963). Perceptual Learning. *Annual Review of Psychology, 14,* 29-56.

Gibson, E.J. (1969). *Principles of perceptual learning.* New York: Appleton-Century-Crofts.

Gibson, E.J., & Walk, R.D. The effect of prolonged exposure to visually presented patterns on learning to discriminate them. *Journal of Comparative and Physiological Psychology, 49,* 239-242.

Gleitman, L.R., Armstrong, S.L. & Gleitman, H. (1983). On doubting the concept "concept". In E.K. Scholnick (Ed.), *New trends in conceptual representation: Challenges to Piaget's theory?* (pp. 87-110). Hillsdale: NJ: Erlbaum.

Gluck, M.A., & Bower, G.H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227-247.

Gluck, M.A., & Myers, C.E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus, 3,* 491-516.

Goldstone, R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52,* 125-157.

Goldstone, R.L. (1998). Perceptual Learning. *Annual Review of Psychology, 49,* 585-612.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects.* New York: The Bobbs-Merrill Co.

Gosselin, F., & Schyns, P.G. (2001). Why do we SLIP to the basic level? Computational constraints and their implementation. *Psychological Review, 108,* 735-758.

Graf, M. (2002). *Form, space and object: Geometrical transformation in object recognition and categorization.* Berlin, Germany: Wissenschaftlicher Verlag Berlin.

Grand, C.S., Close, J., Hale, J, & Honey, R.C. (2007). The role of similarity in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 33,* 64-71.

Grand, C.S., & Honey, R.C. (2008). Solving XOR. *Journal of Experimental Psychology: Animal Behavior Processes, 34,* 486-493.

Grill-Spector, K., & Kanwisher, N. (2005). As Soon as You Know It Is There, You Know What It Is. *Psychological Science, 16,* 152-160.

Gureckis, T.M., & Goldstone, R.L. (2008). The effect of internal structure of categories on perception. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1876-1881). Washington, D.C.: Cognitive Science Society.

Gureckis, T.M., & Love, B.C. (2003). Towards a Unified Account of Supervised and Unsupervised Learning. *Journal of Experimental and Theoretical Artificial Intelligence, 15,* 1-24.

Gureckis, T.M., & Love, B.C. (in press). Direct Associations or Internal Transformations? Exploring the Mechanisms Underlying Sequential Learning Behavior. *Cognitive Science.*

Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts and D. Shanks (Eds.), *Knowledge, Concepts and Categories* (pp. 43-92). Hove, Psychology Press: MIT Press.

Hahn, U., Chater, N., & Richardson, L.B.C. (2003). Similarity as Transformation. *Cognition, 87,* 1-32.

Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape similarity judgments. *Psychological Science, 20,* 447-461.

Hall, G. (1991). *Perceptual and associative learning.* New York: Oxford University Press.

Hampton, J.A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior, 18,* 441-461.

Hampton, J.A. (1981). An investigation of the nature of abstract concepts. *Memory & Cognition, 9,* 149-156.

Hampton, J.A. (1995). Testing Prototype Theory of Concepts. *Journal of Memory and Language, 34,* 686-708.

Hampton, J.A. (2001). The Role of Similarity in Natural Categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and Categorization* (pp. 13-28). Oxford: Oxford University Press.

Hampton, J.A. (2003). Abstraction and Context in Concept Representation. *Philosophical Transactions of the Royal Society of London, Theme Issue:* 'The abstraction paths: from experience to concept', *358,* 1251-1259.

Handel, S., & Imai, S. (1972). The free classification of analyzable and unanalyzable stimuli. *Perception & Psychophysics, 12,* 108-116.

Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition.* New York: Cambridge University Press.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), *Knowledge, Concepts, and Categories* (pp. 7-41). Psychology Press.

Heit, E. (2001). Background knowledge and models of categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and Categorization* (pp. 155-178). Oxford: Oxford University Press.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D.L. Medin (Ed.), *Psychology of Learning and Motivation, Vol. 39* (pp. 163-199). Academic Press.

Herrnstein, R.J. (1979). Acquisition, generalization, and discrimination reversal of a natural concept. *Journal of Experimental Psychology: Animal Behavior Processes, 5,* 118-129.

Herrnstein, R.J. (1985). Riddles of natural categorization. *Philosophical Transactions of the Royal Academy, London, B, 308,* 129-144.

Herrnstein, R.J., & Loveland, D.H. (1964). Complex Visual Concept in the Pigeon. *Science, 146,* 549-551.

Herrnstein, R.J., Loveland, D.H., & Cable, C. (1976). Natural Concepts in Pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 2,* 285-311.

Herrnstein, R.J. (1990). Levels of stimulus control. *Cognition, 37,* 133-166.

Hockema, S.A., Blair, M.R., & Goldstone, R.L. (2005). Differentiation for novel dimensions. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 953-958). Hillsdale, NJ: Erlbaum.

Homa, D., & Cultice, J.C. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 10,* 83-94.

Hodder, K.I., George, D.N., Killcross, A.S., & Honey, R.C. (2003). Representational blending in human conditional learning: Implications for associative theory. *Quarterly Journal of Experimental Psychology, 56,* 223-238.

Honey, R.C., & Bateson, P. (1996). Stimulus comparison and perceptual learning: Further evidence and evaluation from an imprinting procedure. *Quarterly Journal of Experimental Psychology, 49B,* 259-269.

Honey, R.C., Bateson, P., & Horn, G. (1994). The role of stimulus comparison in perceptual learning: An investigation with the domestic chick. *Quarterly Journal of Experimental Psychology, 47B,* 83-103.

Honey, R.C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes, 15,* 338-346.

Honey, R.C., & Ward-Robinson, J. (2001). Transfer between contextual conditional discriminations: An examination of how stimulus conjunctions are represented. *Journal of Experimental Psychology: Animal Behavior Processes, 27*, 196-205.

Honey, R.C., & Ward-Robinson, J. (2002). Acquired equivalence and distinctiveness of cues: I. Exploring a neural network approach. *Journal of Experimental Psychology: Animal Behavior Processes, 28,* 378-387.

Honey, R.C., & Watt, A. (1998). Acquired Relational Equivalence: Implications for the Nature of Associative Structures. *Journal of Experimental Psychology: Animal Behavior Processes, 24,* 325-334.

Honey, R.C., & Watt, A. (1999). Acquired Relational Equivalence Between Contexts and Features. *Journal of Experimental Psychology: Animal Behavior Processes, 25,* 324-333.

Imai, S., & Garner, W.R. (1965). Discriminability and preference for attributes in free and constrained classification. *Journal of Experimental Psychology, 69*, 596-608.

Japkowicz, N. (2001). Supervised Versus Unsupervised Binary-Learning by Feedforward Neural Networks. *Machine Learning, 42,* 97-122.

Japkowicz, N., Myers, C., & Gluck, M.A. (1995). A Novelty Detection Approach to Classification. In *Proceedings of the Fourteenth Joint Conference on Artificial Intelligence* (pp. 518-523). Montreal, CA.

Jitsumori, M. (1993). Category discrimination of artificial polymorphous stimuli based on feature learning. *Journal of Experimental Psychology: Animal Behavior Processes, 19,* 244-254.

Jitsumori, M. (1994). Discrimination of artificial polymorphous categories by rhesus monkeys (*Macaca mulatta*). *Quarterly Journal of Experimental Psychology, 47B,* 371-386.

Johnson-Laird, P.N. (1983). *Mental models.* Cambridge, MA: Harvard University Press.

Jones, G.V. (1983). Identifying basic categories. *Psychological Bulletin, 94,* 423-428.

Kaplan, A.S., & Murphy, G.L. (1999). The Acquisition of Category Structure in Unsupervised Learning. *Memory & Cognition, 27*, 699-712.

Kaplan, A.S., & Murphy, G.L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 829-846.

Katz, J. (1972). *Semantic Theory.* New York: Harper & Row.

Katz, J. & Fodor, J.A. (1963). The structure of a semantic theory. *Language, 39,* 170-210.

Keil, F.C. (1989). *Concepts, kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kemler, D.G. (1984). The effect of intention on what concepts are acquired. *Journal of Verbal Learning and Verbal Behavior, 23,* 734-759.

Kemler, D.G., & Smith, L.B. (1979). Accessing similarity and dimensional relations: Effects of integrality and separability on the discovery of complex concepts. *Journal of Experimental Psychology: General, 108,* 133-150.

Komatsu, L.K. (1992). Recent Views of Conceptual Structure. *Psychological Bulletin, 112,* 500-526.

Krzanowski, W.J., & Marriott, F.H.C. (1995). *Multivariate analysis, Part 2: Classification, covariance structures and repeated measurements.* Arnold: London.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Kurtz, K.J. (1986). Category-based similarity. *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Lakoff, G. (1972). Hedges: A study in meaning criteria and the logic of fuzzy concepts. In P.M. Peranteau, J.N. Levi, & G.C. Phares (Eds.), P*apers from the eighth regional meeting.* Chicago Linguistics Society (pp. 183-228). Chicago: Chicago Linguistics Society.

Lakoff, G. (1987a). Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 63-100). New York: Cambridge University Press.

Lakoff, G. (1987b). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General, 124,* 161-180.

Lamberts, K. (2000). Information-accumulation theory of speeded classification. *Psychological Review, 107*, 227-260.

Lamberts, K. (2002). Feature sampling in categorization and recognition of objects. *Quarterly Journal of Experimental Psychology, 55A,* 141-154.

Lassaline, M.E., & Murphy, G.L. (1996). Induction and category coherence. *Psychonomic Bulletin & Review, 3,* 95-99.

Lassaline, M.E., Wisniewski, E.J., & Medin, D.L. (1992). Basic level in artificial and natural categories: Are all basic levels created equal? In B. Burns (Ed.), *Percepts, Concepts and Categories: The Representation and Processing of Information. Advances in psychology, Vol. 93* (pp. 327-378). Amsterdam: Elsevier.

Lavis, Y., & Mitchell, C. (2006). Effects of preexposure on stimulus discrimination: An investigation of the mechanisms responsible for human perceptual learning. *Quarterly Journal of Experimental Psychology, 59,* 2083-2101.

Lawrence, D.H. (1949). Acquired distinctiveness of cues. I. Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology, 39,* 770-784.

Le Pelley, M.E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *Quarterly Journal of Experimental Psychology, 57B,* 193-243.

Lea, S.E.G. (1984). In what sense do pigeons learn concepts? In H.T. Roitblat, T.G. Bever, & H.S. Terrace (Eds.), *Animal cognition* (pp. 263-276). Hillsdale, NJ: Lawrence Erlbaum Associates Inc..

Lea, S.E.G., & Harrison, S.N. (1978). Discrimination of polymorphous stimulus sets by pigeons. *Quarterly Journal of Experimental Psychology, 30,* 521-537.

Lea, S.E.G., & Wills, A.J. (2008). Use of multiple dimensions in learned discriminations. *Comparative Cognition & Behavior Reviews, 3,* 115-133.

Lea, S.E.G., Wills, A.J., & Ryan, C.M.E. (2006). Why are artificial polymorphous concepts so hard for birds to learn? *Quarterly Journal of Experimental Psychology, 59,* 251-267.

Lewandowsky, S., Roberts, L., & Yang, L. (2006). Knowledge partitioning in categorization: Boundary conditions. *Memory & Cognition, 34,* 1676-1688.

Liberman, A.M., Harris, K.S., Eimas, P.D., Lisker, L., & Bastian, J. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 61,* 379-388.

Livingston, K.R., Andrews, J.K., & Harnad, S. (1998). Categorical Perception Effects Induced by Category Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 732-753.

López, A., Atran, S., Coley, J.D., Medin, D.L., & Smith, E.E. (1997). The Tree of Life: Universal and Cultural Features of Folkbiological Taxonomies and Inductions, *Cognitive Psychology, 32,* 251-295.

Love, B.C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 10,* 190-197.

Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review, 111,* 309-332.

Lubow, R.E. (1989). *Latent Inhibition and Conditioned Attention Theory.* New York: Cambridge University Press.

Malt, B.C. (1990). Features ad beliefs in the mental representation of categories. *Journal of Memory and Language, 29,* 289-315.

Malt, B.C. (1995). Category Coherence in Cross-Cultural Perspective. *Cognitive Psychology, 29,* 85-148.

Malt, B.C., & Sloman, S.A. (2007). Category essence or essentially pragmatic? Creator's intention in naming and what's really what. *Cognition, 105,* 615-648.

Malt, B.C. & Smith, E.E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior, 23,* 250-269.

Mareschal, D., & Quinn, P.C. (2001). Categorization in infancy. *Trends in Cognitive Sciences, 5,* 443-450.

Markman, A.B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25,* 431-467.

Marsh, H., & MacDonald, S. (2008). The use of perceptual features in categorization by orangutans (*Pongo abelli*). *Animal Cognition, 11,* 569-585.

McClelland, J.L., & Rumelhart, D.E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159-197.

McCloskey, M.E., & Glucksberg, S. (1978). Natural categories: Well defined or fuzzy sets? *Memory and Cognition, 6,* 462-472.

McLaren, I.P.L, & Mackintosh, N.J. (2000). Associative learning and elemental representations. I: A theory and its application to latent inhibition and perceptual learning. *Animal Learning & Behavior, 26,* 211-246.

McLaren, I.P.L., & Mackintosh, N.J. (2002). Associative learning and elemental representation: II. Generalization and discrimination. *Animal Learning & Behavior, 30,* 177-200.

McNamara, T.P., & Sternberg, R.J. (1983). Mental models of word meaning. *Journal of Verbal Learning and Verbal Behavior, 22,* 449-474.

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254-278.

Medin, D.L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). New York: Cambridge University Press.

Medin, D.L., & Ross, B.H. (1997). *Cognitive Psychology* (2nd ed.). Fort Worth: Harcourt Brace.

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

Medin, D.L., & Schwanenflugel, P.J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 355-368.

Medin, D.L., & Smith, E.E. (1984). Concepts and concept formation. *Annual Review of Psychology, 35,* 113-138.

Medin, D.L. & Wattenmaker, W.D. (1987). Category cohesiveness, theories, and cognitive archaeology. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 25-62). New York: Cambridge University Press.

Medin, D.L., Wattenmaker, W.D., & Hampson, S.E. (1987). Family resemblance, conceptual cohesiveness and category construction. *Cognitive Psychology, 19*, 242-279.

Mercado, E., III., Orduña, I., & Nowak, J.N. (2005). Auditory categorization of complex sounds by rats. *Journal of Comparative Psychology, 119,* 90-98.

Mervis, C.B. & Rosch, E.R. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-115.

Milton, F., & Wills, A.J. (2004). The influence of stimulus properties on category construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 407-415.

Milton, F., Longmore, C.A., & Wills, A.J. (2008). Processes of overall similarity sorting in free classification. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 676-692.

Mishkin, M., Prockop, E.S., & Rosvold, H.E. (1962). One-trial object-discrimination learning in monkeys with frontal lesions. *Journal of Comparative and Physiological Psychology, 55,* 178-181.

Morgan, M.J. (2005). The visual computation of 2-D area by human observers. *Vision Research, 45,* 2564-2570.

Morgan, M.J., Fitch, M.D., Holman, M.D., & Lea, S.E.G. (1976). Pigeons learn the concept of an 'A'. *Perception, 5,* 57-66.

Mundy, M.E., Honey, R.C., & Dwyer, D.M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33,* 124-138.

Mundy, M.E., Honey, R.C., & Dwyer, D.M. (2009). Superior discrimination between similar stimuli after simultaneous exposure. *Quarterly Journal of Experimental Psychology, 62,* 18-25.

Murai, C., Tomonaga, M., Kamegai, K., Terazawa, N., & Yamaguchi, K.M. (2004). Do infant Japanese macaques (*Macaca fuscata*) categorize objects without specific training? *Primates, 45,* 1-6.

Murphy, G.L. (2002). *The big book of concepts.* MIT Press: Cambridge, USA.

Murphy, G.L., & Allopenna, P.D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

Murphy, G.L. & Brownell, H.H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 70-84.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Murphy, G.L. & Smith, E.E. (1982). Basic level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior, 21*, 1-20.

Murphy, G.L. & Wisniewski, E.J. (1989). Categorizing objects in isolation and in scenes: What a superordinate is good for. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 572-586.

Nakagawa, E. (1986). Overtraining, extinction, and shift learning in concurrent discriminations. *Quarterly Journal of Experimental Psychology, 38B*, 313-326.

Neisser, U. (1987). From direct perception to conceptual structure. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 11-24). Cambridge, England: Cambridge University Press.

Nelson, K. (1974). Concept, word, and sentence: Interrelations in acquisition and development. *Psychological Review, 81,* 267-285.

Newell, F.N., & Bülthoff, H.H. (2002). Categorical perception of familiar objects. *Cognition, 85,* 113-143.

Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 104-114.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R.M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 700-708.

Nosofsky, R.M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, memory and cognition, 14*, 54-65.

Nosofsky, R.M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics, 45,* 279-290.

Nosofsky, R.M. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19*, 131-150.

Nosofsky, R.M., Clark, S.E., & Shin, H.J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282-304.

Nosofsky, R.M., Johansen, M.K. (2000). Exemplar-based accounts of multiple-system phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7,* 375-402.

Nosofsky, R.M., & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

Oden, D.L., Thompson, R.K.R., & Premack, D. (1988). Spontaneous transfer of matching by infant chimpanzees (*Pan troglodytes*). *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 140-145.

Ogden, C.K., & Richards, I.A. (1956). *The meaning of meaning*. New York: Harcourt, Brace.

Pastore, R.E. (1987). Categorical perception: Some psychophysical models. In S. Harnad (Ed.), *Categorical Perception: The Groundwork of Cognition* (pp. 29-52). New York: Cambridge University Press.

Pavlov, I.P. (1927). *Conditioned reflexes*. London: Oxford University Press.

Pearce, J.M. (1987). A model of stimulus generalization for Pavlovian conditioning. *Psychological Review, 94,* 61-73.

Pearce, J.M. (1988). Stimulus generalization and the acquisition of categories by pigeons. In L. Weiskrantz (Ed.), *Thought without language* (pp. 132-152). Oxford: Oxford University Press.

Pearce, J.M. (1989). The acquisition of an artificial category by pigeons. *Quarterly Journal of Experimental Psychology, 41B,* 381-406.

Pearce, J.M. (1991). The acquisition of abstract and concrete categories by pigeons. In L. Dachowski, & C. Flaherty (Eds.), *Current topics in animal learning: Brain, emotion and, Cognition* (pp. 141-164). New Jersey: L. Erlbaum.

Pearce, J.M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review, 101,* 587-607.

Pearce, J. M. (1997). *Animal Learning and Cognition: An Introduction (2nd Ed.).* East Sussex, UK: Psychology Press Ltd.

Pickering, M., & Chater, N. (1995). Why cognitive science is not formalized folk Psychology. *Minds and Machines, 5,* 309-337.

Pomerantz, J.R., & Kubovy, M. (1986). Theoretical approaches to perceptual organization: Simplicity and likelihood principles. In K.R. Boff, L. Kaufman, & J.P. Thomas (Eds.), *Handbook of perception and human performance, volume II: Cognitive processes and performance* (pp. 1-45). New York: Wiley.

Porter, D., Neuringer, A. (1984). Music discriminations by pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 10,* 138-148.

Posner, M.I., & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353-363.

Pothos, E.M., Bailey, T.M. (2009). Predicting category intuitiveness with the rational model, the simplicity model, and the Generalized Context Model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1062-1080.

Pothos, E.M., & Chater, N. (2001). Categorization by simplicity: a minimum description length approach to unsupervised clustering. In U. Hahn & M. Ramscar (Eds.), *Similarity and Categorization* (pp. 51-72). Oxford: Oxford University Press.

Pothos, E.M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26,* 303-343.

Pothos, E.M., & Close, J. (2008). One or two dimensions in spontaneous classification: A matter of simplicity. *Cognition, 107,* 581-602.

Pothos, E.M., Hahn, U., & Prat-Sala, M. (2009). Similarity chains in the transformational paradigm. *European Journal of Cognitive Psychology, 21,* 1100-1120.

Putnam, H. (1975). The meaning of 'meaning'. In H. Putnam (Ed.), *Mind, Language, and reality: Philosophical papers, vol. 2.* Cambridge, England: Cambridge University Press.

Putnam, H. (1996). Meaning and Reference. In A.P. Martinich (Ed.), *The Philosophy of Language* (pp. 284-291). New York: OUP, Inc.

Premack, D. (1976). *Intelligence in Ape and Man.* Hillsdale, NJ: Erlbaum.

Premack, D. (1983). The codes of Man and Beasts. *Behavioral and Brain Sciences, 6,* 125-137.

Quinlan, R.J., & Rivest, R.L. (1989). Inferring decision trees using the Minimum Description Length Principle. *Information and Computation, 80,* 227-248.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66,* 846-850.

Reed, S.K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Regehr, G., & Brooks, L.R. (1995). Category organization in free classification: the organizing effect of an array of stimuli. *Journal of Experimental Psychology Learning, Memory, & Cognition. 21*, 347-363.

Rescorla, R.A. (1991). Associative relations in instrumental learning: The eighteenth Bartlett memorial lecture. *Quarterly Journal of Experimental Psychology, 43B*, 1-23.

Rescorla, R.A., & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Rips, L.J., Blok, S., & Newman, G. (2006). Tracing the identity of objects. *Psychological Review, 113,* 1-30.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14,* 465-471.

Rissanen, J. (1989). *Stochastic complexity and statistical inquiry.* Singapore: World Scientific.

Rogers, T.T. & Patterson, K. (2007). Object Categorization: Reversals and Explanations of the Basic-Level Advantage. *Journal of Experimental Psychology: General, 136*, 451-469.

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4,* 328-350.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104,* 192-233.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C.B. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, *7*, 573-605.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382-439.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* London: Sage.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533-536.

Ryan, C.M.E. (1982). Concept formation and individual recognition in the domestic chicken. *Behaviour Analysis Letters, 2,* 213-220.

Ryan, C.M.E., & Lea, S.E.G. (1990). Pattern Recognition, Updating, and Filial Imprinting in the Domestic Chicken (*Gallus gallus*). In M.L. Commons, R.J. Herrnstein, S.M. Kosslyn, & D.B. Mumford (Eds.), *Behavioral Approaches to Pattern Recognition and Concept Formation* (pp. 89-110). Hillsdale, NJ: Erlbaum.

Schnur, P., & Lubow, R.E. (1976). Latent inhibition: The effects of ITI and CS intensity during preexposure. *Learning and Motivation, 7,* 540-550.

Schrier, A.M., Angarella, R., & Povar, M.L. (1984). Studies of concept formation by stumptailed monkeys: Concepts humans, monkeys, and letter A. *Journal of Experimental Psychology: Animal Behavior Processes, 10,* 564-584.

Schrier, A.M., & Brady, P.M. (1987). Categorization of natural stimuli by monkeys (Macaca mulatta): Effects of stimulus set size and modification of exemplars. *Journal of Experimental Psychology: Animal Behavior Processes, 13,* 136-143.

Schyns, P.G. (1991). A modular neural network model of concept acquisition. *Cognitive Science, 15,* 461-508.

Seamon, J.G. (1982). Dynamic facial recognition: Examination of a natural phenomenon. *American Journal of Psychology, 95,* 363-381.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

Shin, H.J., & Nosofsky, R.M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General, 121,* 278-304.

Siegel, R.K., & Honig, W.K. (1970). Pigeon concept formation: Successive and simultaneous acquisition. *Journal of the Experimental Analysis of Behavior, 13,* 385-390.

Smith, E.E., & Medin, D.L. (1981). *Categories and concepts.* Cambridge, MA: Harvard University Press.

Smith, J.D., & Kemler Nelson, D.G. (1984). Overall similarity in adults' description: The child in all of us. *Journal of Experimental Psychology: General, 113,* 137-159.

Smith, L.B. & Heise, D. (1992). Perceptual similarity and conceptual structure. In B. Burns (Ed.), *Percepts, concepts, and categories* (pp. 233-272). New York: Elsevier.

Smith, L.B., & Kemler, D.G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology, 24,* 279-298.

Spalding, T.L., & Murphy, G.L. (1996). Effects of background knowledge on category construction. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 525-538.

Spinozzi, G. (1993). Development of spontaneous classificatory behavior in chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology, 107,* 193-200.

Spinozzi, G. (1996). Categorization in monkeys and chimpanzees. *Behavioral Brain Research, 74,* 17-24.

Spinozzi, G., & Natale, F. (1989). Classification. In F. Antinucci (Ed.), *Cognitive structures and development in nonhuman primates* (pp. 163-187). Hillsdale, NJ: Erlbaum.

Spinozzi, G., Natale, F., Langer, J., & Brakke, K. (1999). Spontaneous class grouping behavior by bonobos (*Pan paniscus*) and common chimpanzees (*Pan troglodytes*). *Animal Cognition, 2,* 157-170.

Sutherland, N.S., & Mackintosh, N.J. (1971). *Mechanisms of animal discrimination learning*. New York: Academic Press.

Symonds, M., & Hall, G. (1995). Perceptual learning in favor aversion conditioning: Roles of stimulus comparison and latent inhibition of common stimulus elements. *Learning & Motivation, 26,* 203-219.

Tanaka, J.W. & Taylor, M.E. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 15,* 121-149.

Tenenbaum, J.B., Griffiths, T.L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10,* 309-318.

Thompson, R.K.R., Oden, D.L., & Boysen, S.T. (1997). Language-naive chimpanzees (*Pan troglodytes*) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes, 23,* 31-43.

Thorpe, S., & Imbert, M. (1989). Biological constraints on connectionist modelling. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulié, & L. Steels (Eds.), *Connectionism in perspective* (pp. 63-93). Amsterdam: Elsevier.

Troje, N.F., Huber, L., Loidolt, M., Aust, U., & Fieder, M. (1999). Categorical learning in pigeons: The role of texture and shape in complex static stimuli. *Vision Research, 39,* 353-366.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

Tversky, A., & Gati, I. (1978). Studies of Similarity. In E. Rosch & B.B. Lloyd (Eds.), *Cognition and categorization* (pp. 79-98). Hillsdale: Erlbaum.

Urcuioli, P.J., Zentall, T.R., & DeMarse, T. (1995). Transfer to derived sample-comparison relations by pigeons following many-to-one versus one-to-many matching with identical training relations. *Quarterly Journal of Experimental Psychology*, *48B*, 158-178.

Vaughan, W., & Greene, S.L. (1984). Pigeon visual memory capacity. *Journal of Experimental Psychology: Animal Behavior Processes, 10,* 256-271.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 1: behavioural study. *European Journal of Neuroscience, 11,* 1223-1238.

Vonk, J., & MacDonald, S.E. (2002). Natural concepts in a juvenile gorilla (G*orilla gorilla gorilla*) at three levels of abstraction. *Journal of the Experimental Analysis of Behavior, 78,* 315-332.

Vonk, J., & MacDonald, S.E. (2004). Levels of Abstraction in Orangutan (*Pongo abelii*) Categorization. *Journal of Comparative Psychology, 118,* 3-13.

Ward, T.B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to-separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance, 9,* 103-112.

Ward, T.B., Foley, C.M., & Cole, J. (1986). Classifying multidimensional stimuli: Stimulus, task, and observer factors. *Journal of Experimental Psychology: Human Perception and Performance, 12,* 211-225.

Wasserman, E.A., Kiedinger, R.E., & Bhatt, R.S. (1988). Conceptual behavior in pigeons: Categories, subcategories, and pseudocategories. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 235-246.

Wattenmaker, W.D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17,* 908-923.

Wattenmaker, W.D., Dewey, G.I., Murphy, T.D., & Medin, D.L. (1986). Linear separability and concept learning: Context, relational properties, and naturalness. *Cognitive Psychology*, *18*, 158-194.

Wilson, B., Mackintosh, N.J., & Boakes, R.A. (1985). Transfer of relational rules in matching and oddity learning by pigeons and corvids. *Quarterly Journal of Experimental Psychology, 37B,* 313-332.

Wisniewski, E.J., & Medin, D.L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, *18*, 221-281.

Wittgenstein, L. (1953). *Philosophical Investigations.* New York: Macmillan.

Young, M.E., & Wasserman, E.A. (1997). Entropy Detection by Pigeons: Response to Mixed Visual Displays After Same-Different Discrimination Training. *Journal of Experimental Psychology: Animal Behavior Processes, 23,* 157-170.

Young, M.E., & Wasserman, E.A. (2001). Entropy and variability discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 278-293.

Zaki, S.R., & Homa, D. Concepts and Transformational Knowledge. *Cognitive Psychology, 39,* 69-115.

Zentall, T.R., Steirn, J.N., Sherburne, L.M., & Urcuioli, P.J. (1991). Common coding in pigeons assessed through partial versus total reversals of many-to-one conditional and simple discriminations. *Journal of Experimental Psychology: Animal Behavior Processes, 17, 194-201.*

Appendices

Appendix 1

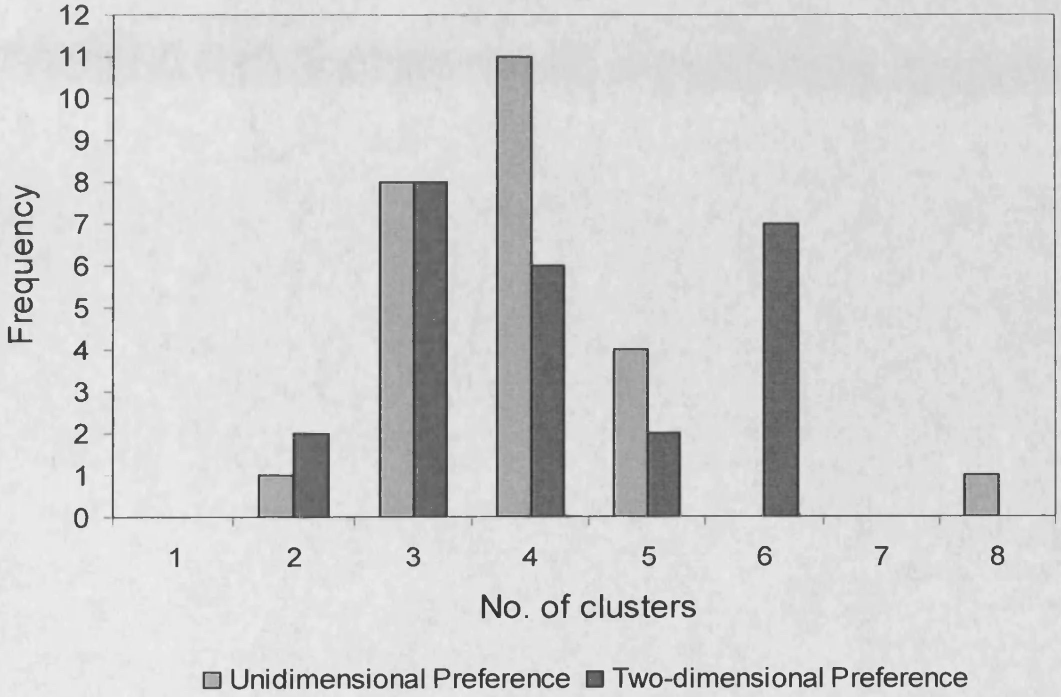

Figure 28. The frequency with which participants produced classifications based on a specific number of clusters in Experiment 1.

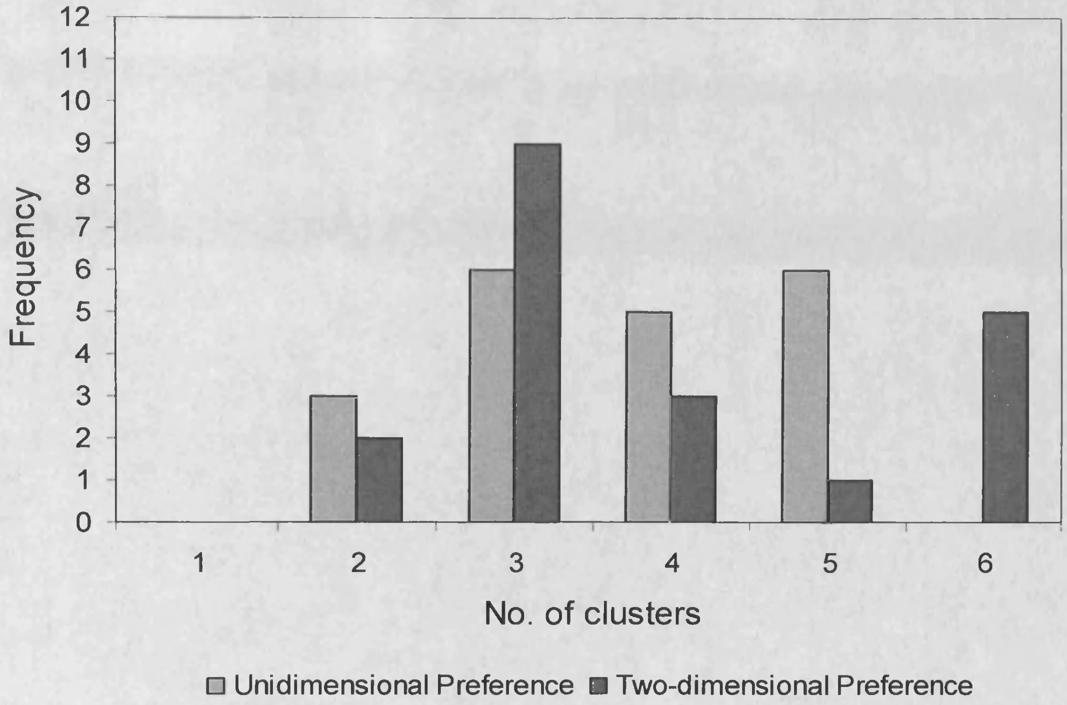

Figure 29. The frequency with which participants produced classifications based on a specific number of clusters in Experiment 2.

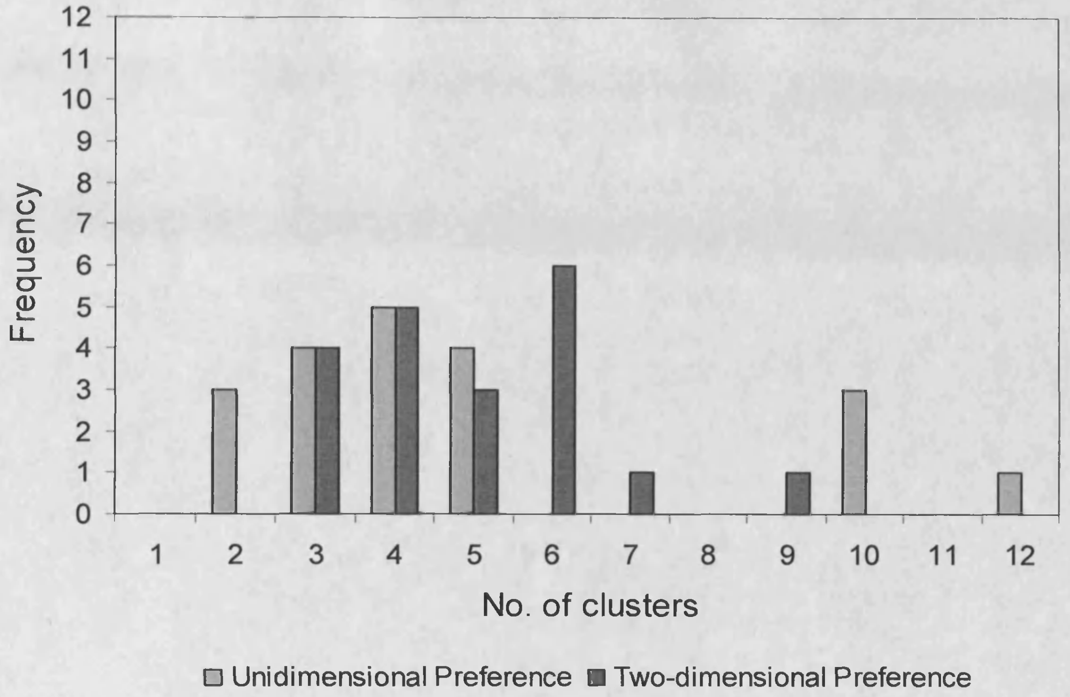

Figure 30. The frequency with which participants produced classifications based on a specific number of clusters in Experiment 3.

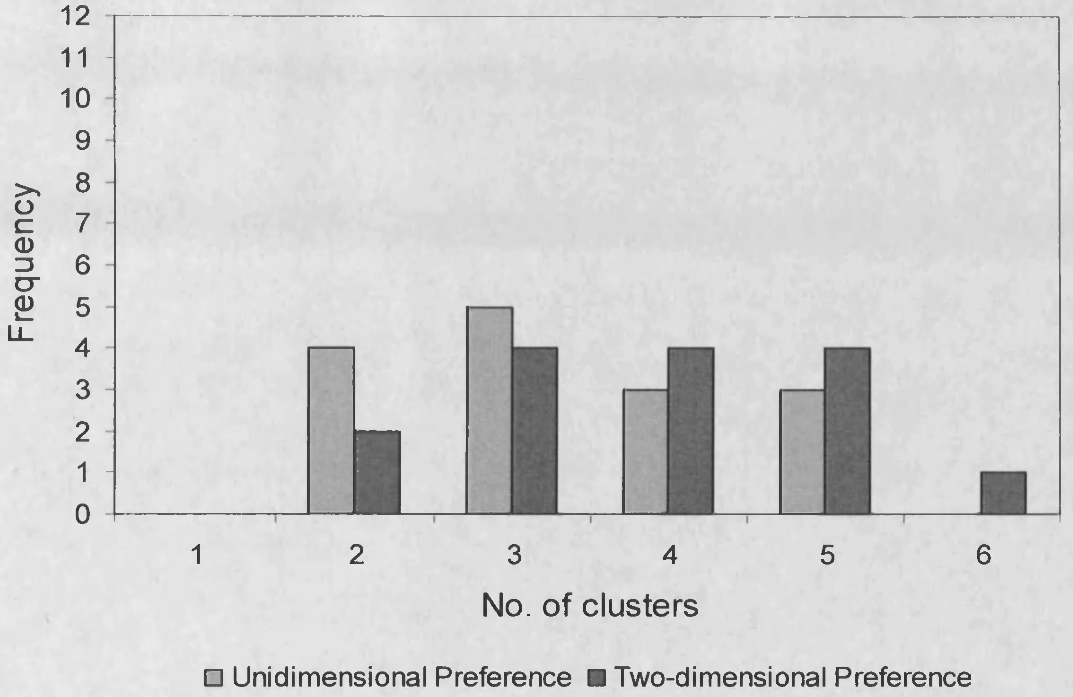

Figure 31. The frequency with which participants produced classifications based on a specific number of clusters in Experiment 4.

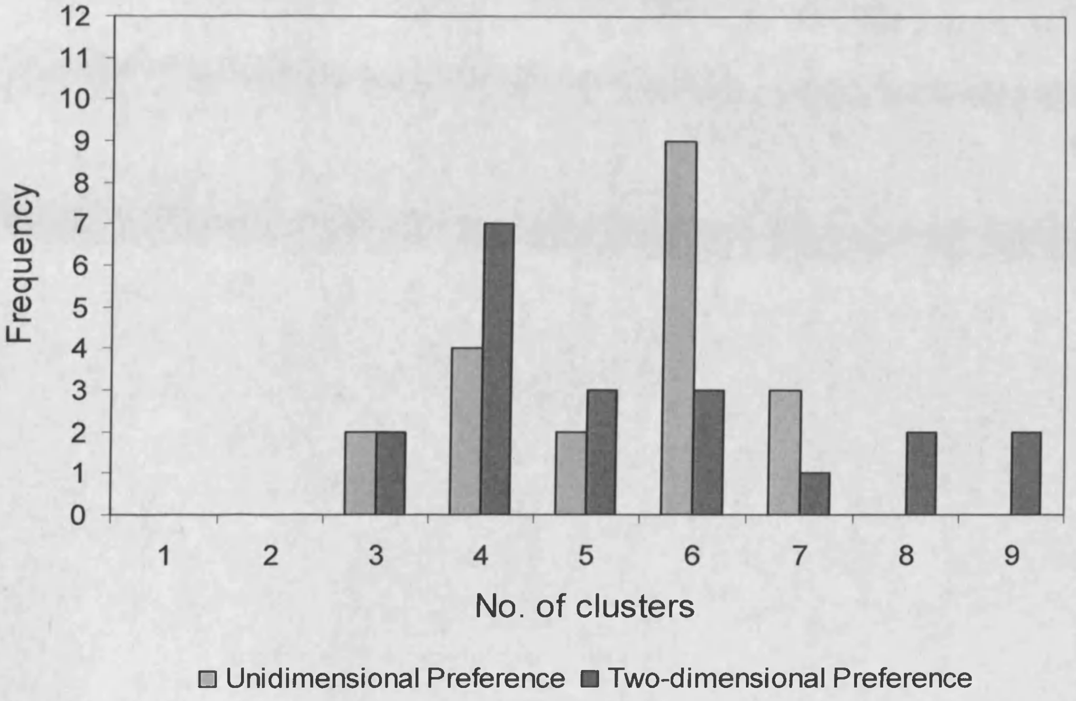

Figure 32. The frequency with which participants produced classifications based on a specific number of clusters in Experiment 5.

Appendix 2

Table 8. *Mean number of food well entries during PCS and CS periods across the three days of conditioning, split by condition.*

	Experiment 9					
	Day 1		Day 2		Day 3	
Condition	PCS	CS	PCS	CS	PCS	CS
Baseline	66.25	70.63	61.75	74.63	57.38	64.75
Surprise	59.38	64.75	53.75	64.63	48.88	59.13
Sys_trans	51.38	62.25	47.50	48.75	40.13	52.88
Scram_trans	56.00	67.75	52.63	68.38	43.63	67.63
	Experiment 10					
	Day 1		Day 2		Day 3	
Condition	PCS	CS	PCS	CS	PCS	CS
Baseline	63.63	75.13	55.75	66.25	48.13	59.13
Sys_trans	56.63	58.00	39.75	50.50	39.50	59.00
	Experiment 11					
	Day 1		Day 2		Day 3	
Condition	PCS	CS	PCS	CS	PCS	CS
Baseline	61.88	71.13	40.50	54.75	42.13	66.00
Sys_trans	53.88	67.50	48.75	56.75	47.25	70.50